Same Neurons, Different Languages: Probing Morphosyntax in Multilingual Pre-trained Models

Anonymous ACL submission

Abstract

The success of multilingual pre-trained models in transferring knowledge cross-lingually is underpinned by their ability to learn representations shared by multiple languages even in absence of any explicit supervision. However, it remains unclear *how*. In this work, we conjecture that multilingual pre-trained models can derive language-universal abstractions about grammar. In particular, we investigate whether morphosyntactic information is encoded in the same subset of neurons in different languages. We conduct the first large-scale empirical study over 43 typologically diverse languages and 14 morphosyntactic categories with a state-of-the-art neuron-level probe. Our findings show that the cross-lingual overlap between neurons is significant, but its extent may vary across categories and depends on language proximity and pretraining data size.

1 Introduction

Massively multilingual pre-trained models (Devlin et al., 2019; Conneau et al., 2020; Liu et al., 2020; Xue et al., 2021, *inter alia*) display an impressive ability to transfer knowledge between languages and perform zero-shot inference (Pires et al., 2019; Wu and Dredze, 2019). Nevertheless, it remains unclear how pre-trained models learn multilingual representations despite the lack of an explicit signal through parallel texts. While some speculate that overlap in sub-words plays a key role in this process (Wu and Dredze, 2019; Cao et al., 2020), Artetxe et al. (2020) provide contrary evidence.

In this work, we conjecture that multilingual representations are facilitated by the fact that—in addition to lexical alignment (Pires et al., 2019; Vulić et al., 2020)—the neurons dedicated to specific morphosyntactic categories (such as gender for nouns and mood for verbs) are shared across languages.[1] We validate this hypothesis empirically by probing 3 multilingual pre-trained models, m-BERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) base and large, for morphosyntactic information in 43 typologically diverse languages from Universal Dependencies (Nivre et al., 2017). In particular, we use the state-of-the-art intrinsic probe of Anonymous (2021) inspired by Torroba Hennigen et al. (2020), which can identify small subsets of neurons in a representation that jointly encode morphosyntactic information. We collect compelling evidence that, while trained independently, these probes find similar neuron subsets in multiple languages.

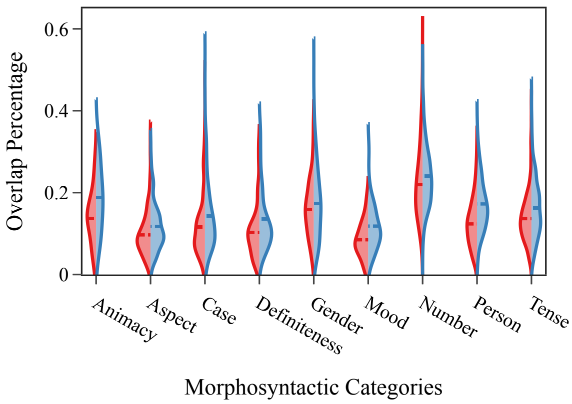


Figure 1: Percentages of neurons most associated with a particular morphosyntactic category that overlap between pairs of languages. Colours in the plot refer to 2 models: m-BERT (red) and XLM-R-base (blue).

Moreover, we discover that language pairs with high proximity (in the same genus or with similar typological features) and with large amounts of pretraining data tend to exhibit more overlap. In addition, more neurons are shared in models with *less* parameters and for morphosyntactic categories with a small inventory of possible values.

2 Background

First, we must determine which neurons in a model representation encode a particular linguistic prop-

[1] Concurrent work by Antverg and Belinkov (2021) suggests a similar hypothesis based on smaller-scale experiments.

erty, which is known as intrinsic probing (Dalvi et al., 2019). In particular, we adopt the methodology of Torroba Hennigen et al. (2020), where we aim to find a subset of k neurons $C^\star \subseteq D = \{1, \ldots, d\}$, where d is the total number of dimensions in the representation being probed, that jointly maximise some performance measure $\mathcal{S}$

$$C^\star = \underset{\substack{C \subseteq D,\\ |C|=k}}{\operatorname{argmax}}\, \mathcal{S}(C) \tag{1}$$

Following Torroba Hennigen et al. (2020), we choose the log-likelihood of a probe evaluated on held out data as $\mathcal{S}$, and solve the objective in Eq. (1) by greedy selection.

Even with greedy selection, however, the objective in Eq. (1) is intractable. This is because this procedure would require training a separate probe for every different subset of dimensions under consideration, which means $\frac{n!}{k!\,(d-k)!}$ times. To address this, we resort to the probe of Anonymous (2021), which can be trained once and yields a parameterisation that works well regardless of which subset of features is being evaluated. Furthermore, Anonymous (2021) find that this approach outperforms previous intrinsic probes from Torroba Hennigen et al. (2020) and Dalvi et al. (2019).

Anonymous (2021) achieve this by sampling random dimensions during training as a regularisation. More formally, let Π be the inventory of values that some morphosyntactic category can take in a particular language, for example $\Pi = \{\textsc{Feminine}, \textsc{Masculine}, \textsc{Neutral}\}$ for grammatical gender in Russian. Moreover, let $\mathcal{D} = \{(\pi^{(n)}, \boldsymbol{h}^{(n)})\}_{n=1}^{N}$ be a dataset of labelled embeddings such that $\pi^{(n)} \in \Pi$ and $\boldsymbol{h}^{(n)} \in \mathbb{R}^d$, where d is the dimensionality of the representation being considered, e.g., $d = 768$ for m-BERT. Anonymous (2021) observe that marginalising over subsets of informative neurons C, one can derive an expression for the log-likelihood of a neural model with parameters $\boldsymbol{\theta}$

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log p_{\boldsymbol{\theta}}(\pi^{(n)} \mid \boldsymbol{h}^{(n)}) \tag{2}$$

$$= \sum_{n=1}^{N} \log \sum_{C \subseteq D} p_{\boldsymbol{\theta}}\left(\pi^{(n)} \mid \boldsymbol{h}^{(n)}, C\right) p(C)$$

where we opt for an (uninformative) uniform prior $p(C)$, similarly to Anonymous (2021). This objective is still intractable due to the sum over 2^d subsets of dimensions. Hence, we optimise the variational lower bound (ELBo) of Eq. (2) instead. In particular, we introduce a variational distribution $q_{\boldsymbol{\phi}}(C)$ over subsets of neurons

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log \sum_{C \subseteq D} p_{\boldsymbol{\theta}}\left(\pi^{(n)}, C \mid \boldsymbol{h}^{(n)}\right) \tag{3}$$

$$\geq \sum_{n=1}^{N} \left(\mathop{\mathbb{E}}_{C \sim q_{\boldsymbol{\phi}}} \left[\log p_{\boldsymbol{\theta}}(\pi^{(n)}, C \mid \boldsymbol{h}^{(n)}) \right] + \mathcal{H}(q_{\boldsymbol{\phi}}) \right)$$

where $\mathcal{H}(\cdot)$ stands for the entropy of a distribution. The full derivation of Eq. (3) is provided in App. A. For this paper, we chose $q_{\boldsymbol{\phi}}(\cdot)$ to correspond to a Poisson sampling scheme (Hájek, 1964), where subsets of dimensions are sampled by subjecting each dimension to an independent Bernoulli trial. The variational parameters ϕ correspond to the unnormalised probability of sampling a particular dimension.[2]

3 Method

In our work, we learn distinct intrinsic probes for 43 languages and for 14 categories as described in §2. Then we assess whether the neuron overlaps between languages are statistically significant.

Data. We select 43 treebanks from Universal Dependencies 2.1 (UD; Nivre et al., 2017), which contain sentences annotated with morphosyntactic information in a wide array of languages. Afterwards, we compute contextual representations for every individual word in the treebanks using multilingual BERT (m-BERT) and the base and large versions of XLM-RoBERTa (XLM-R-base and XLM-R-large). We then associate each word with its parts of speech and morphosyntactic features, which are mapped to the UniMorph schema (Kirov et al., 2018).[3] The selected treebanks include all languages supported by both BERT and XLM-R which are available in UD.

Rather than adopting the default UD splits, we re-split word representations based on lemmata ending up with disjoint vocabularies for the train, development, and test set. This prevents a probe from achieving high performance by sheer memorising. Moreover, for every category–language pair (e.g., mood–Czech), we discard any lemma with fewer than 20 tokens in its split.

[2] We opt for this sampling scheme as Anonymous (2021) found that it is more computationally efficient than conditional Poisson while achieving a comparable performance.

[3] We use the converter from McCarthy et al. (2018).

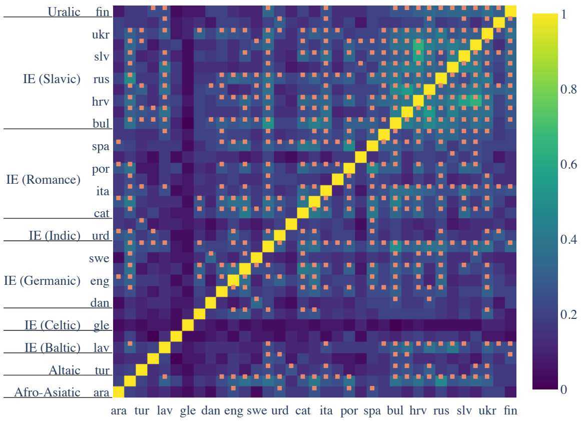


Figure 2: The percentage overlap between the top-50 most informative number dimensions in m-BERT for number. Statistically significant overlap is marked with an orange square.

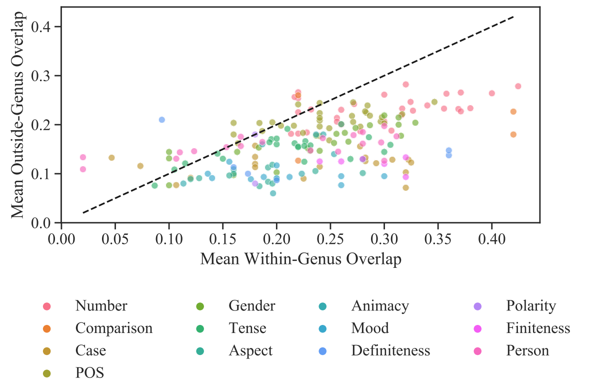


Figure 3: Mean percentage of neuron overlap in XLM-R-base with languages either within or outside the same genus for each morphosyntactic category.

Experimental Setup. We first train a probe for each morphosyntactic category–language combination with the objective in Eq. (3). In line with established practices in probing, we parameterise $p_{\boldsymbol{\theta}}(\cdot)$ as a linear layer followed by a softmax. Afterwards, we identify the top-k most informative neurons in the last layer of m-BERT, XLM-R-base, and XLM-R-large. Specifically, following Torroba Hennigen et al. (2020), we use the log-likelihood of the probe on the test set as our greedy selection criterion. Thus, we single out 50 dimensions for each combination of morphosyntactic category and language.

Next, we measure the pairwise overlap in the top-k most informative dimensions between all pair of languages where a morphosyntactic category is expressed. This results in matrices such as Fig. 2, where the pair-wise percentages of overlapping dimensions is visualised as a heat map.

Statistical Significance. Suppose that two languages have $m = \{1, \ldots, k\}$ overlapping neurons when considering the top-k selected neurons for each of them. To determine whether such overlap is statistically significant, we compute the probability of an overlap of *at least* m neurons under the null hypothesis that the sets of neurons are sampled independently at random. We estimate these probabilities with a permutation test. In this paper, we set a threshold of $\alpha = 0.05$ for significance. Finally, we use Holm-Bonferroni (Holm, 1979) family-wise error correction as detailed in App. C. Hence, our threshold is appropriately adjusted for multiple comparisons, which makes incorrectly rejecting the null hypothesis more likely. For instance, in Fig. 2, statistically significant pairs are marked with an orange square.

4 Results

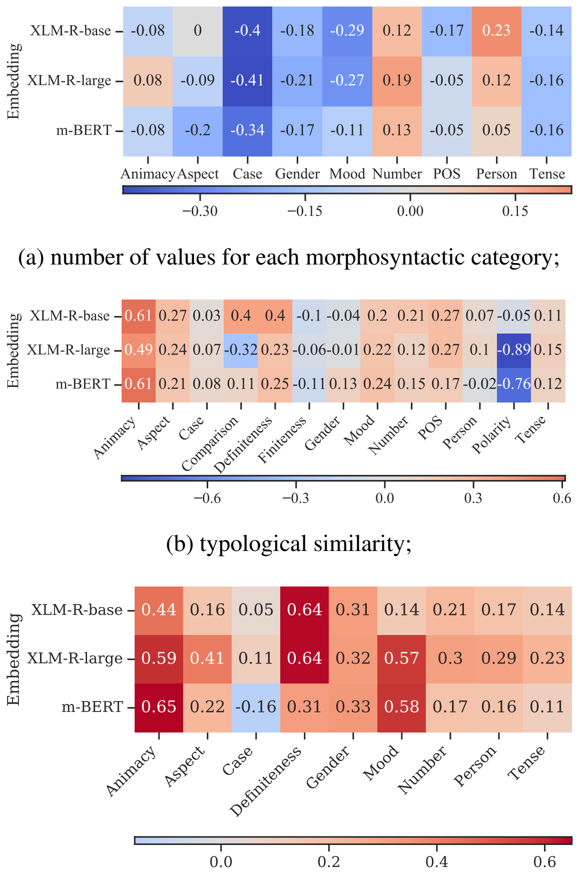
We begin by analysing our claim that multilingual pre-trained models develop a cross-lingually entangled notion of morphosyntax. The matrices of pairwise overlaps for each of the 14 categories, such as Fig. 2 for number, are reported in App. E. We condense these results in two distinct ways. First, we report the cross-lingual distribution for each category in in Fig. 1 for m-BERT and XLM-R-base.[4] Moreover, we calculate how many overlaps are statistically significant out of the total number of pairwise comparisons in Tab. 1. From these figures, it emerges that around 20% of neurons among the top-50 most informative ones overlap on average, but the number of statistically significant ones may vary dramatically across categories.

Morphosyntactic Categories. Based on Tab. 1, significant overlap is particularly accentuated in specific categories, such as comparison, polarity, and number. However, neurons for other categories such as mood, aspect, and case are shared by only a handful of language pairs despite the high number of comparisons. This finding may be partially explained by the different number of values each category can take. Hence, we test whether there is a correlation between this number and average cross-lingual overlap in Fig. 4a. As expected, we generally find negative correlation coefficients—prominent exceptions being number and person. As the inventory of values of a category grows, cross-lingual alignment becomes harder.

Language Proximity. Moreover, we investigate whether language proximity, in terms of both lan-

[4] An equivalent plot comparing XLM-R-base and XLM-R-large is available in Fig. 5.

Figure 4: Spearman's correlation, for a given model and morphological category, between the cross-lingual average percentage of overlapping neurons and:



(c) language model training data size.

guage family and typological features, bears any relationship with the neuron overlap for any particular pair. In Fig. 3, we plot pairwise similarities with languages within the same genus (e.g., Baltic) against those outside. From the distribution of the dots, we can extrapolate than sharing of neurons is more likely to occur between languages in the same genus. This is further corroborated by the language groupings emerging in the matrices of App. E.

In Fig. 4b, we also measure the correlation between neuron overlap and similarity of syntactic typological features based on Littell et al. (2017). While correlation coefficients are mostly positive (with the exception of polarity), we remark that the patterns is strongly influenced by whether a category is typical for a specific genus. For instance, correlation is highest for animacy, a category almost exclusive to Slavic languages in our sample.

Pre-trained models. Afterwards, we determine whether the 3 models under consideration reveal different patterns. Comparing m-BERT and XLM-R-base in Fig. 1, we find that, on average, XLM-R-base tends to share more neurons when encoding particular morphosyntactic attributes. Moreover, comparing XLM-R-base to XLM-R-large in Fig. 5 suggests that more neurons are shared in the former than in the latter. Altogether, these results seem to suggest that the presence of additional training data engenders cross-lingual entanglement, but increasing model size incentivises morphosyntactic information to be allocated to different subsets of neurons. We conjecture that this may be best viewed from the lens of compression: If model size is a bottleneck, then, to attain good performance across many languages, a model must learn cross-lingual abstractions that can be reused.

Pre-training data size. Finally, we assess the effect of pre-training data size[5] for neuron overlap in every language. According to Fig. 4c, their correlation is very high. We explain this phenomenon with the fact that more data yields higher-quality (and hence, more entangled) multilingual representations.

5 Conclusions

In this paper, we hypothesise that the ability of multilingual models to generalise across languages results from cross-lingually entangled representation, where the same subsets of neurons encode universal morphosyntactic information. We validate this claim with a large-scale empirical study on 43 typologically diverse languages and 3 models, namely m-BERT, XLM-R-base, and XLM-R-large. Based on our empirical results, we conclude that the overlap is statistically significant for a considerable amount of language pairs. However, the extent of the overlap varies remarkably across morphosyntactic categories and tends to be lower for categories with large inventories of possible values. Moreover, we found that neuron subsets are shared mostly between languages in the same genus or with similar typological features. Finally, we discover that the overlap of each language grows proportionally to its pre-training data size, but it also decreases in larger model architectures.

In future work, artificially encouraging a tighter neuron overlap might facilitate zero-shot cross-lingual inference to low-resource and typologically distant languages (Zhao et al., 2021).

[5]We rely on the CC-100 statistics reported by Conneau et al. (2020) for XLM-R and on the Wikipedia dataset's size with TensorFlow datasets (Abadi et al., 2015) for m-BERT.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems.

Anonymous. 2021. A latent-variable model for intrinsic probing.

Omer Antverg and Yonatan Belinkov. 2021. On the pitfalls of analyzing individual neurons in language models.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. 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.

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

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.

Jaroslav Hájek. 1964. Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, 35(4):1491–1523.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. UniMorph 2.0: Universal morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Arya D. McCarthy, Miikka Silfverberg, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2018. Marrying universal dependencies and universal morphology. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 91–101, Brussels, Belgium. Association for Computational Linguistics.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Eckhard Bick, Victoria Bobicev, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Aljoscha Burchardt, Marie Candito, Gauthier Caron, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Savas Cetin, Fabricio Chalub, Jinho Choi, Silvie Cinková, Çağrı Çöltekin, Miriam Connor, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Ali Elkahky, Tomaž Erjavec, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier

Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Nizar Habash, Jan Hajič, Jan Hajič jr., Linh Hà Mỹ, Kim Harris, Dag Haug, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Radu Ion, Elena Irimia, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Hiroshi Kanayama, Jenna Kanerva, Tolga Kayadelen, Václava Kettnerová, Jesse Kirchner, Natalia Kotsyba, Simon Krek, Veronika Laippala, Lorenzo Lambertino, Tatiana Lando, John Lee, Phương Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Gustavo Mendonça, Niko Miekka, Anna Missilä, Cătălin Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Shinsuke Mori, Bohdan Moskalevskyi, Kadri Muischnek, Kaili Müürisep, Pinkey Nainwani, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lương Nguyễn Thị, Huyền Nguyễn Thị Minh, Vitaly Nikolaev, Hanna Nurmi, Stina Ojala, Petya Osenova, Robert Östling, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Emily Pitler, Barbara Plank, Martin Popel, Lauma Pretkalniņa, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Larissa Rinaldi, Laura Rituma, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Benoît Sagot, Shadi Saleh, Tanja Samardžić, Manuela Sanguinetti, Baiba Saulīte, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Takaaki Tanaka, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Lars Wallin, Jonathan North Washington, Mats Wirén, Tak-sum Wong, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, and Hanzhi Zhu. 2017. Universal dependencies 2.1. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

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.

Lucas Torroba Hennigen, Adina Williams, and Ryan Cotterell. 2020. Intrinsic probing through dimension selection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 197–216, Online. Association for Computational Linguistics.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. 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.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Wei Zhao, Steffen Eger, Johannes Bjerva, and Isabelle Augenstein. 2021. Inducing language-agnostic multilingual representations. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 229–240, Online. Association for Computational Linguistics.

A Variational Lower Bound

The derivation of the variational lower bound is shown below:

$$
\begin{aligned}
&\sum_{n=1}^{N} \log \sum_{C \subseteq D} p_{\boldsymbol{\theta}}(\pi^{(n)}, C \mid \boldsymbol{h}^{(n)}) \\
&= \sum_{n=1}^{N} \log \sum_{C \subseteq D} q_{\boldsymbol{\phi}}(C) \frac{p_{\boldsymbol{\theta}}(\pi^{(n)}, C \mid \boldsymbol{h}^{(n)})}{q_{\boldsymbol{\phi}}(C)} \\
&= \sum_{n=1}^{N} \log \mathop{\mathbb{E}}_{C \sim q_{\boldsymbol{\phi}}} \left[\frac{p_{\boldsymbol{\theta}}(\pi^{(n)}, C \mid \boldsymbol{h}^{(n)})}{q_{\boldsymbol{\phi}}(C)} \right] \\
&\geq \sum_{n=1}^{N} \mathop{\mathbb{E}}_{C \sim q_{\boldsymbol{\phi}}} \left[\log \frac{p_{\boldsymbol{\theta}}(\pi^{(n)}, C \mid \boldsymbol{h}^{(n)})}{q_{\boldsymbol{\phi}}(C)} \right] \\
&= \sum_{n=1}^{N} \left(\mathop{\mathbb{E}}_{C \sim q_{\boldsymbol{\phi}}} \left[\log p_{\boldsymbol{\theta}}(\pi^{(n)}, C \mid \boldsymbol{h}^{(n)}) \right] + \mathcal{H}(q) \right)
\end{aligned}
$$

B Probed Property–Language Pairs

- **afr (Afrikaans)**: Part of Speech, Number, Tense
- **ara (Arabic)**: Gender, Voice, Mood, Part of Speech, Aspect, Person, Number, Case, Definiteness
- **bel (Berlarusian)**: Part of Speech, Tense, Number, Aspect, Finiteness, Voice, Gender, Animacy, Case, Person
- **bul (Bulgarian)**: Part of Speech, Definiteness, Gender, Number, Mood, Tense, Person, Voice, Comparison
- **cat (Catalan)**: Gender, Number, Part of Speech, Tense, Mood, Person, Aspect
- **ces (Czech)**: Part of Speech, Number, Case, Comparison, Gender, Mood, Person, Tense, Aspect, Polarity, Animacy, Possession, Voice
- **dan (Danish)**: Part of Speech, Number, Gender, Definiteness, Voice, Tense, Mood, Comparison
- **deu (German)**: Part of Speech, Case, Number, Tense, Person, Comparison
- **ell (Greek)**: Part of Speech, Case, Gender, Number, Finiteness, Person, Tense, Aspect, Mood, Voice, Comparison
- **eng (English)**: Part of Speech, Number, Tense, Case, Comparison
- **est (Estonian)**: Part of Speech, Mood, Finiteness, Tense, Voice, Number, Person, Case
- **eus (Basque)**: Part of Speech, Case, Animacy, Definiteness, Number, Argument Marking, Aspect, Comparison
- **fas (Persian)**: Number, Part of Speech, Tense, Person, Mood, Comparison
- **fin (Finnish)**: Part of Speech, Case, Number, Mood, Person, Voice, Tense, Possession, Comparison
- **fra (French)**: Part of Speech, Number, Gender, Tense, Mood, Person, Polarity, Aspect
- **gle (Irish)**: Tense, Mood, Part of Speech, Number, Person, Gender, Case
- **glg (Galician)**: Part of Speech
- **heb (Hebrew)**: Part of Speech, Number, Tense, Person, Voice
- **hin (Hindi)**: Person, Case, Part of Speech, Number, Gender, Voice, Aspect, Mood, Finiteness, Politeness
- **hrv (Croatian)**: Case, Gender, Number, Part of Speech, Person, Finiteness, Mood, Tense, Animacy, Definiteness, Comparison, Voice
- **ita (Italian)**: Part of Speech, Number, Gender, Person, Mood, Tense, Aspect
- **jpn (Japanese)**: Part of Speech
- **lat (Latin)**: Part of Speech, Number, Gender, Case, Tense, Person, Mood, Aspect, Comparison
- **lav (Latvian)**: Part of Speech, Case, Number, Tense, Mood, Person, Gender, Definiteness, Aspect, Comparison, Voice
- **lit (Lithuanian)**: Tense, Voice, Number, Part of Speech, Finiteness, Mood, Polarity, Person, Gender, Case, Definiteness
- **mar (Marathi)**: Case, Gender, Number, Part of Speech, Person, Aspect, Tense, Finiteness
- **nld (Dutch)**: Person, Part of Speech, Number, Gender, Finiteness, Tense, Case, Comparison
- **pol (Polish)**: Part of Speech, Case, Number, Animacy, Gender, Aspect, Tense, Person, Polarity, Voice
- **por (Portuguese)**: Part of Speech, Person, Mood, Number, Tense, Gender, Aspect
- **ron (Romanian)**: Definiteness, Number, Part of Speech, Person, Aspect, Mood, Case, Gender, Tense
- **rus (Russian)**: Part of Speech, Case, Gender, Number, Animacy, Tense, Finiteness, Aspect, Person, Voice, Comparison
- **slk (Slovak)**: Part of Speech, Gender, Case, Number, Aspect, Polarity, Tense, Voice, Animacy, Finiteness, Person, Mood, Comparison
- **slv (Slovenian)**: Number, Gender, Part of Speech, Case, Mood, Person, Finiteness, Aspect, Animacy, Definiteness, Comparison

- **spa (Spanish)**: Part of Speech, Tense, Aspect, Mood, Number, Person, Gender
- **srp (Serbian)**: Number, Part of Speech, Gender, Case, Person, Tense, Definiteness, Animacy, Comparison
- **swe (Swedish)**: Part of Speech, Gender, Number, Definiteness, Case, Tense, Mood, Voice, Comparison
- **tam (Tamil)**: Part of Speech, Number, Gender, Case, Person, Polarity, Finiteness, Tense
- **tur (Turkish)**: Case, Number, Part of Speech, Aspect, Person, Mood, Tense, Polarity, Possession, Politeness
- **ukr (Ukrainian)**: Case, Number, Part of Speech, Gender, Tense, Animacy, Person, Aspect, Voice, Comparison
- **urd (Urdu)**: Case, Number, Part of Speech, Person, Finiteness, Voice, Mood, Politeness, Aspect
- **vie (Vietnamese)**: Part of Speech
- **zho (Chinese)**: Part of Speech

C Family-wise Error Correction

The method for estimating statistical significance works for any pair of languages; however, as we are performing multiple comparisons, we should expect the null hypothesis to be incorrectly rejected $100 \times \alpha$ percent of the time. To circumvent this problem, we resort to Holm-Bonferroni (Holm, 1979) family-wise error correction.

In particular, the tests are ordered in an ascending order by means of their p-values. The test with the smallest probability undergoes the Holm–Bonferroni correction

$$p_{\text{HB}} = (n - i + 1)p \tag{4}$$

where n denotes the number of conducted tests. If already the first test is not significant, the procedure stops, otherwise the test with the second smallest p-value is corrected for a family of $n - 1$ tests. The procedure stops either at the first non-significant test or after iterating though all p-values. This sequential approach guarantees that probability that we incorrectly reject *one or more* of our hypotheses is at most α.

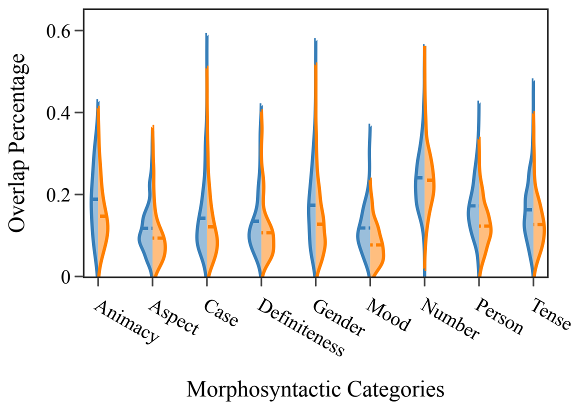
D Overlap Rates

Tab. 1 depicts the proportion of neuron overlap for different attributes and embeddings.

	m-BERT	XLM-R-base	XLM-R-large	Total
Definiteness	0.11	0.22	0.13	45
Comparison	0.20	0.90	0.50	10
Possession	0.00	0.00	0.00	1
Aspect	0.03	0.10	0.09	153
Polarity	0.33	0.67	0.33	3
Number	0.40	0.51	0.74	666
Animacy	0.14	0.57	0.32	28
Mood	0.00	0.07	0.05	105
Gender	0.15	0.32	0.19	378
Person	0.08	0.25	0.13	276
POS	0.04	0.27	0.70	861
Case	0.10	0.18	0.17	300
Tense	0.08	0.23	0.12	325
Finiteness	0.09	0.18	0.09	45

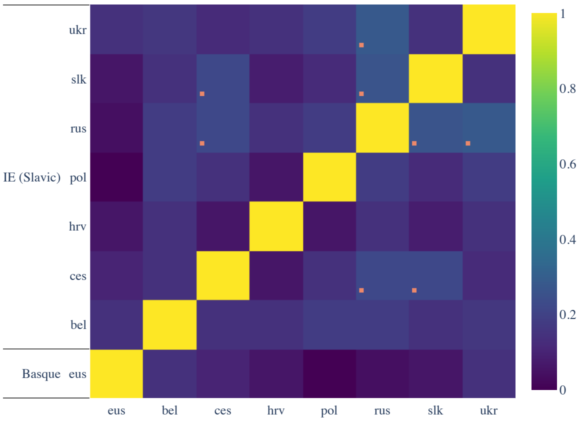
Table 1: Proportion of language pairs with statistically significant overlap in the top-50 neurons for an attribute (after Holm–Bonferroni (Holm, 1979) correction). We compute these proportions for each model we consider. The final column reports the total number of pairwise comparisons.

Figure 5: Percentages of neurons most associated with a particular morphosyntactic category that overlap between pairs of languages. Colours in the plot refer to 2 models: XLM-R-base (blue) and XLM-R-large (orange).

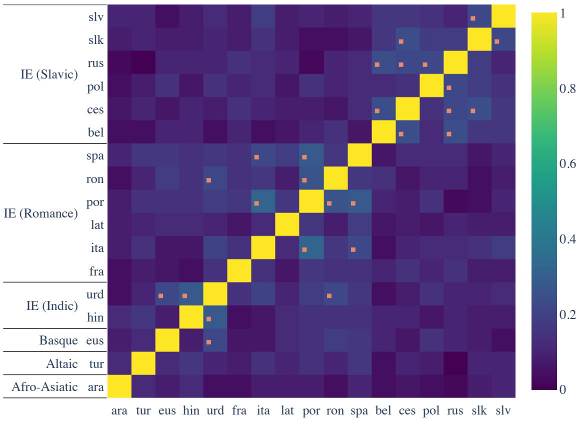


E Pairwise Overlap by Morphosyntactic Category

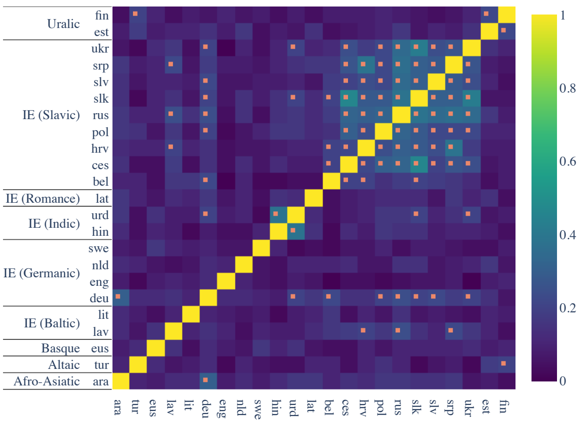
Figure 6: The percentage overlap between the top-50 most informative dimensions in a randomly selected language model for each of the morphosyntactic categories. Statistically significant overlap is marked with an orange square.



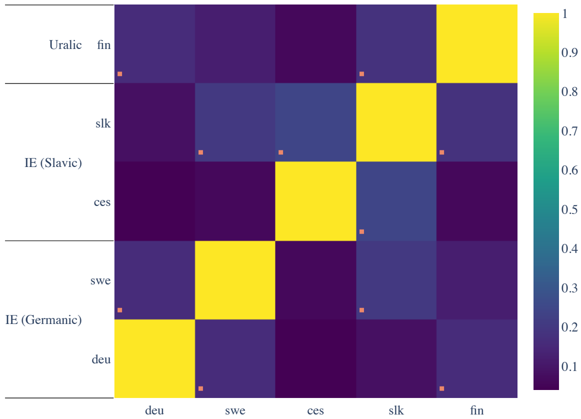
(a) Animacy–m-BERT



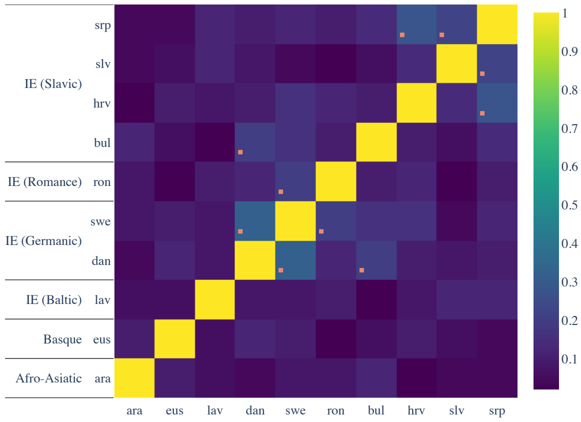
(b) Aspect–XLM-R-base



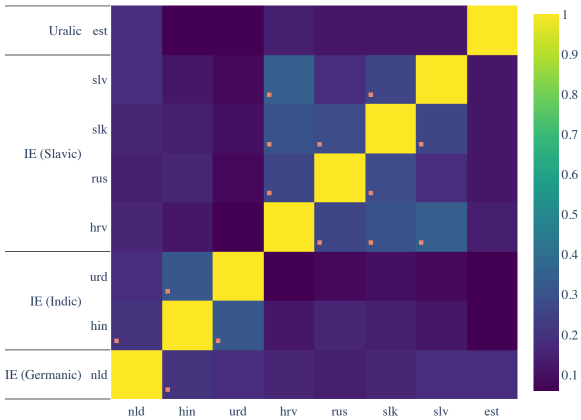
(c) Case–XLM-R-large



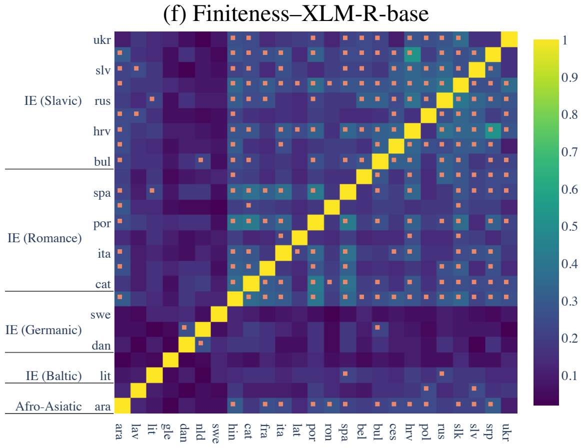
(d) Comparison–XLM-R-large



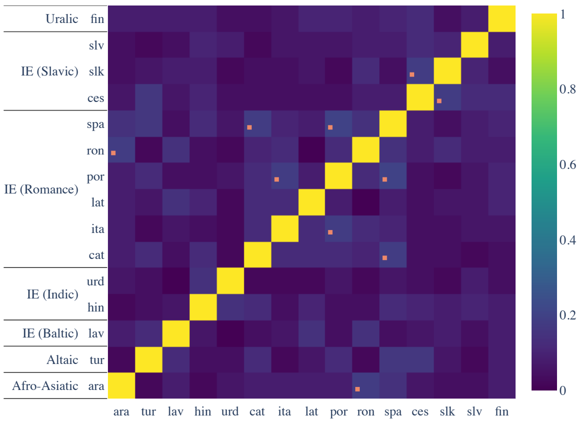
(e) Definiteness–m-BERT



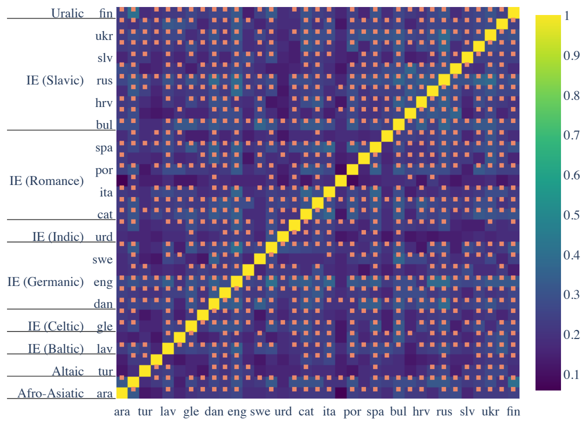
(f) Finiteness–XLM-R-base



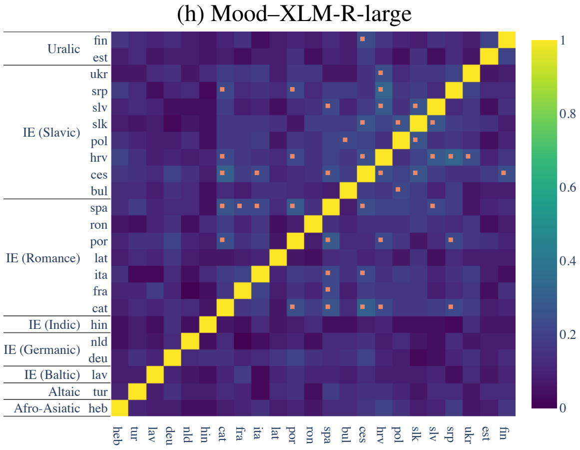
(g) Gender–XLM-R-base



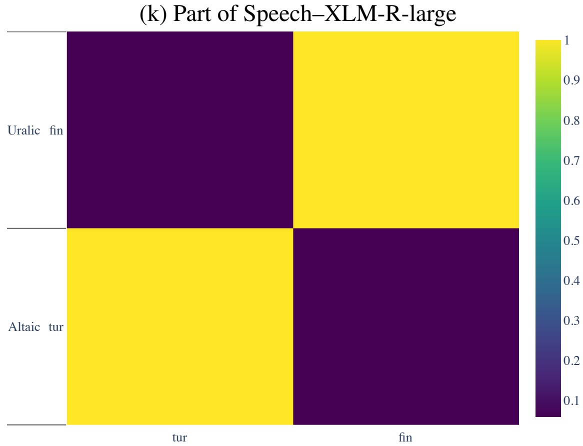
(h) Mood–XLM-R-large



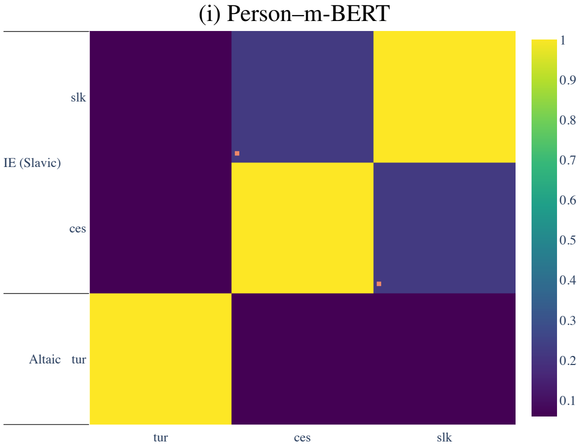
(k) Part of Speech–XLM-R-large



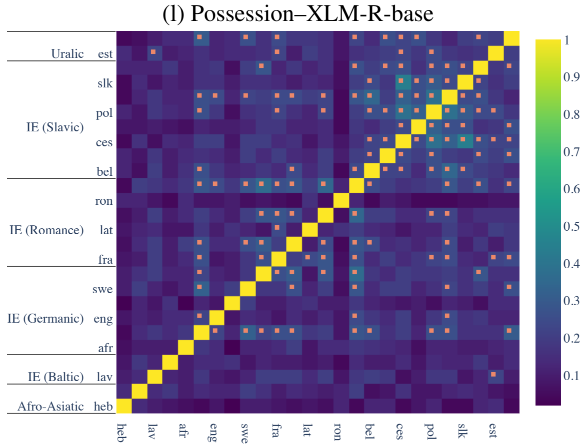
(i) Person–m-BERT



(l) Possession–XLM-R-base



(j) Polarity–XLM-R-large



(m) Tense–XLM-R-base