Universal and Independent: Multilingual Probing Framework for Exhaustive Model Interpretation and Evaluation

Oleg Serikov[‡♠♡], **Vitaly Protasov**[‡], **Ekaterina Voloshina**[$], **Viktoria Knyazkova**[♡], **Tatiana Shavrina**[‡$]

[‡] Artificial Intelligence Research Institute, [$] SberDevices,
[♡] HSE University, [♠] DeepPavlov lab, MIPT

Abstract

Linguistic analysis of language models is one of the ways to explain and describe their reasoning, weaknesses, and limitations. In the probing part of the model interpretability research, studies concern individual languages as well as individual linguistic structures. The question arises: are the detected regularities linguistically coherent, or on the contrary, do they dissonate at the typological scale? Moreover, the majority of studies address the inherent set of languages and linguistic structures, leaving the actual typological diversity knowledge out of scope. In this paper, we present and apply the GUI-assisted framework allowing us to easily probe a massive number of languages for all the morphosyntactic features present in the Universal Dependencies data. We show that reflecting the anglo-centric trend in NLP over the past years, most of the regularities revealed in the mBERT model are typical for the western-European languages. Our framework can be integrated with the existing probing toolboxes, model cards, and leaderboards, allowing practitioners to use and share their standard probing methods to interpret multilingual models. Thus we propose a toolkit to systematize the multilingual flaws in multilingual models, providing a reproducible experimental setup for 104 languages and 80 morphosyntactic features. GitHub

1 Introduction

Probing methods shed light on the black box of the neural models in unearthing the linguistic features encoded in them. Probing sets a standard setup with various internal representations from the model and uses an auxiliary classifier to predict linguistic information captured in the representation.

As probing research results have come up with contradictory results on different languages and language models, there appears to be a methodological need for a meta-study of the accumulated knowledge and a need to standardize the experimental setup. At the same time, the fixation of the setup and hyperparameters should allow the reproduction of a wide range of experiments, such as multilingual probing, like X-Probe (Ravishankar et al., 2019a) and Linspector (Sahin et al., 2020), layer-wise probing (Fayyaz et al., 2021), chronological probing (Voloshina et al., 2022).

Often, data for probing experiments is based on already known competition data, benchmarks, and gold standards. To obtain consistent results, such data must be high-quality, manually validated, and carefully include multiple languages. For this reason, in this work, we use the Universal Dependencies data (de Marneffe et al., 2021) as a source of multilingual data with a validated and standardized complete morphological and syntactic annotation, which will allow us to accumulate the assimilation of specific linguistic phenomena in many languages at once. Probing these languages on the respective annotated linguistic categories would reveal how models seize the typological proximity of languages.

Therefore, the general probing methodology should include (according to Conneau and Kiela (2018)) 1) a fixed set of evaluations based on what appears to be community consensus; 2) a fixed evaluation pipeline with standard hyperparameters; 3) a straightforward Python interface.

This paper aims to extrapolate the multilingual linguistic diversity on the proven and tested SentEval-like methodology.

We state our contribution as follows:

- We develop a framework for exhaustive multilingual probing of the language models, with a complete enumeration of all grammatical characteristics and all languages available in Universal Dependencies while maintaining the standard SentEval format.
- We provide a setup for better and explanatory aggregation and exploration of the massive

Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 441 - 456

probing results with thousands of experiments for each model.

- We illustrate the possibilities of the framework on the example of the mBERT model, demonstrating new insights and reassuring the results of previous studies on narrower data.

Performing probing studies on such a large scale addresses the vision outlined in Nichols (2007) and contribute to a new dimension to linguistic typology research, as the revealed structures are encapsulated in tools and data inseparably tied to nowadays linguistic nature. Our framework provides users from different fields, including linguists, with a new point of view on the typological proximity of languages and categories.

2 Related Work

Different attempts were made to interpret behavior and hidden learned representation of language models. For example, Hoover et al. (2020) investigated the attention-heads of the BERT model on word tokens connectivity level. Wallace et al. (2019) presented an interpretation framework where they improved a visual component of the model prediction process on several NLP tasks for the end-user.

Flourishing after the ACL debates on semantic parsing[1], the probing methodology has developed its own model interpretation tools. Thus, **SentEval framework** (Conneau and Kiela, 2018) includes various types of linguistically-motivated tasks: surface tasks probe for sentence length (SentLen) and for the presence of words in the sentence (WC); syntactic tasks test for sensitivity to word order (BShift), the depth of the syntactic tree (TreeDepth) and the sequence of top-level constituents in the syntax tree (TopConst); semantic tasks check for the tense (Tense), the subject (resp. direct object) number in the main clause (SubjNum, resp. ObjNum), the sensitivity to random replacement of a noun/verb (SOMO) and the random swapping of coordinated clausal conjuncts (CoordInv).

Linspector (Şahin et al., 2019) includes 15 probing tasks for 24 languages by taking morphosyntactic language properties into account, including case, verb mood, and tense, syntactic correctness, and the semantic impossibility of an example. While lacking the simplicity of the SentEval approach, the framework provides both a linguistically-grounded and multilingual setup. We are significantly expanding both the list of languages and properties being examined.

Probe-X (Ravishankar et al., 2019b) has expanded SentEval setup with 5 additional languages, while **NeuroX framework** (Dalvi et al., 2019) also introduced novelty, but proposed to enrich the methodology to allow for cross-model analysis of the results, supporting neuron-level inspection.

2.1 Probing Critique

We would state a few problems why some of the probing practices are methodologically problematic.

First, the probing interpretation result can differ from paper to paper, creating various conclusions from different authors. While Jawahar et al. (2019) achieves from 69.5-96.2% accuracy on the SentLen SentEval probing task (BERT model), they state that this info is somehow represented at the bottom layers. The work (Ravishankar et al., 2019b) achieves 38-51% accuracy on SentLen (RNN encoder) and states that "recurrent encoders show solid performance on certain tasks, such as sentence length." This drastic difference in result interpretation (“somehow” vs. “extremely strong”) leads to misrepresenting the factual results. Conflicting evidence within the field of BERTology can be found in Rogers et al. (2020), see Sec 3.1 and 4.3.

Secondly, the results on similar tasks can be obtained with unstable success if the hyperparameters are not fixed or exhaustively described: for example, study (Jawahar et al., 2019) finds that "BERT’s intermediate layers encode a rich hierarchy of linguistic information, with surface features at the bottom, syntactic features in the middle and semantic features at the top," while the work by Tikhonova et al. (2022) on mBERT shows, that the model does not learn the linguistic information. More meta-research is needed to explore the contradictory results obtained by the community.

2.2 Task Representation

In the survey of post-hoc language model interpretation (Madsen et al., 2021), the linguistic information-based tasks fall into the groups of the highest abstraction and the top-informativeness of properties used. This group of projects includes tasks based on the various theoretical language levels: from part-of-speech tagging to discourse.

[1] https://aclanthology.org/volumes/W14-24/

Languages While the most tasks are English-based, there appear the non-English monolingual frameworks: French-based probing (Merlo, 2019), Russian-based SentEval (Mikhailov et al., 2021), Chinese word masking probing (Cui et al., 2021). The multilingual benchmarks have paved the way for multilingual probing studies by collecting the necessary data.

Linguistic features Most language-based tasks tend to be based on morphology or syntax, deriving from SentEval methodology. Thus, higher-level tasks can concentrate both on monolingual discourse evaluation (Koto et al., 2021) (mostly English-based by now), as well as the multilingual discursive probing based on the conversion of the existing multilingual benchmarks (Kurfalı and Östling, 2021) (XNLI, XQUAD).

3 Framework Design

This section describes the probing framework and the experimental setup part.

The main goal is to probe how well a model assimilates language constructions during training. For the framework, we want to form an end-to-end solution that can be applied to different models, work on diverse data, and simplify the process of getting insights from the results.

Based on that, the challenges we have are the following:

1. The data we use in the training and evaluation parts must be in the standard format no matter what language we deal with.
2. The probing process should be universal for different models. Based on it, we also need to collect detailed results for further analysis.
3. Since we aim to work with diverse data, we should contain instruments to simplify the process of getting insights from the results. If we do not handle this problem, we can have bunches of results that would be difficult to interpret and provide findings for.

Thus, we can represent our framework as a tool with different instruments. The first one is aimed at pre-processing data for probing, which is commonly a classification task. The second one is a probing engine supporting popular probing techniques such as diagnostic classification. And the last one is a visualization instrument which should ease the process of interpreting the findings.

3.1 SentEval Format Converter

We found the SentEval format to be generally good and universal in the data composition for classification tasks. Since we have such a vast resource as Universal Dependencies for different languages, we can transform the data into the SentEval format and compose different classification tasks based on the language categories we can get.

UD annotation consists of several parts: lemmas, parts of speech, morphological features, and universal dependencies relations. The converter to SentEval format is focused on morphological features. As Table 1 illustrates, morphological categories are written in the sixth column with their category values separated by the equals sign, for example, in *Number=Sing*, *Number* is a category and *Sing* is a category value. It took us 8 hours to process by the SentEval converter on 96 CPUs for absolutely all archives.

For each morphological category found in a given file, the converter generates a new file in SentEval format according to the following steps:

```
Data: CONLLU files or a directory to such
      files for one language
Result: a file in SentEval format
read files;
find all morphological categories;
foreach categories do
    foreach sentences do
        if category is in sentence then
            get a category value
    end
    stratified split on three samples;
    write to a file
end
```

Algorithm 1: The conversion process

If split UD data into train, validation, and test sets, we do not change this split. In other cases, we split data into three sets, so the distribution of category values in the original text will be kept in each set.

If a sentence contains several words with the same morphological categories, the closest to the sentence node word is taken, preventing the one sentence from being repeated several times. Table 1 depicts the example of *Tense* category, the value of word *stopped* will be taken, as it is the root of the sentence.

```
# sent_id = weblog-typepad.com_ripples_20040407125600_ENG_20040407_125600-0055
# text = That too was stopped.
1       That    that    PRON    DT      Number=Sing|PronType=Dem        4       nsubj:pass      4:nsubj:pass    _
2       too     too     ADV     RB      _       4       advmod  4:advmod        _
3       was     be      AUX     VBD     Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin   4       aux:pass        4:aux:pass      _
4       stopped stop    VERB    VBN     Tense=Past|VerbForm=Part|Voice=Pass     0       root    0:root  SpaceAfter=No
5       .       .       PUNCT   .       _       4       punct   4:punct _
```

Figure 1: The example of UD annotation

Format	Data entry
Conll-U	# sent_id = weblog-typepad.com_ripples_20040407125600_ENG_20040407_125 # text = That too was stopped. 1. **That** that PRON DT Number=Sing|PronType=Dem 4 nsubj:pass 4:nsubj:pass _ 2. **too** too ADV RB _ 4 advmod 4:advmod _ 3. **was** be AUX VBD Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin 4 aux:pass 4:aux:pass _ 4. **stopped** stop VERB VBN Tense=Past|VerbForm=Part|Voice=Pass 0 root 0:root SpaceAfter=No 5. . . PUNCT . _ 4 punct 4:punct _
SentEval	tr Past That too was stopped .

Table 1: Example of CONLL-U format and its conversion to SentEval: Tense classification, train set.

3.2 Multilingual Data

We take 289 repositories, including the data of 172 languages available at the GitHub of Universal Dependencies , updated in May 2022.[2]

While parsing files, we face several problems inherited from UD. 71 of the repositories do not contain any CONLLU files. Three Japanese repositories and Korean and Frisian Dutch repositories contain different annotations from standard UD annotations. The data from 16 repositories (Akkadian, Cantonese, Chinese (2), German, Japanese, Hindi, Irish, Kangri, Maltese, Neapolitan, South Levantine Arabic, Swedish Sign language, Swiss German, Old Turkish, Tagalog) do not contain morphological annotation. Also, some repositories include correctly annotated data but are not suitable for classification problems because all the examples contain only one value of all the categories, for example, only examples with class *Plural* are left for the category Number (Cantonese, Chukchi, Frisian Dutch, Hindi English, Japanese, Kangri, Khunsari, Makurap, Maltese, Nayini, Neapolitan, Old Turkish, Soi, South Levantine Arabic, Swedish Sign Language, Swiss German, Telugu, Vietnamese).

After filtering, we have data from 104 languages from 194 repositories (see Appendix A.1). From the typological point of view, these languages belong to 20 language families, and the Basque language is an isolate. Although almost half of the languages are from the Indo-European family, the data include several under-studied language families. Many of the languages in our data are endangered or even extinct. The UD data is distributed based on Creative Commons and GNU-based licenses, varying from language to language[3]. Extracting the tasks for every grammatical category results in 1927 probing datasets.

3.3 Probing Engine

3.3.1 Encoders

In the experiments, we consider the layers of encoder-based models and their ability to acquire language data and perform well on probing tasks. Using the output of the model's layers, we can get contextualized token embeddings for elements of the input text. For that reason, we can consider several options for embedding aggregation: **CLS** where the text is presented as the embedding from "[CLS] "token, **SUM** and **AVG** where the sentence vector is a sum or average of embeddings of all text tokens.

3.3.2 Classifiers and metrics

After the embeddings are obtained, we train a simple classification model based on the encoder layers' representation and task data labels. We consider linear (Logistic Regression) and non-linear (MLP) classifiers. As the metrics for performance evaluation, we use *accuracy* score and weighted F_1 score in case of unbalanced classes.

[2] https://github.com/UniversalDependencies

[3] https://lindat.mff.cuni.cz/repository/xmlui/page/licence-UD-2.1

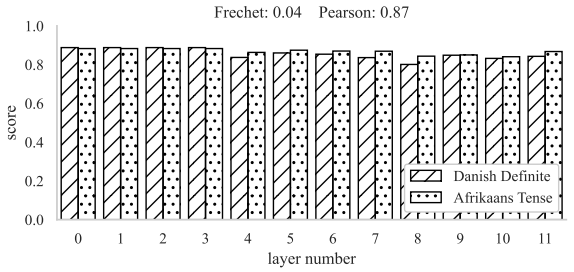

Figure 2: An example of δ_F and r scores calculation between the two probing experiments

3.4 Aggregation

The engine is meant to produce probes of a particular category in a particular language. We provide additional grouping and visualization tools to allow for meaningful interpretation of such large result sets. They are meant to highlight similar experiments and visualize them on the world map.

The default configuration follows the classical probing experiments and uses layers' numbers as X axes. Yet the more novel setup can be chosen, e.g. treating the $<$ language, category $>$ features pairs as X-axis instead.

The defined atomic experimental axis allows to characterize larger groups of experiments via their pooled value (such as mean-pooled by categories value in Figure 6), or even cluster them (e.g., using pairwise experiments similarity as in Figure 3).

3.4.1 Similarity Metrics

We support two metrics of scoring the experiments' pair-wise similarity. Both of them are calculated for the experiment results curves. [4] *Frechet distance* (δ_F) provides a natural way to compare curves taking into account both the similarity of curves' shapes and their absolute positioning on the chart. Unlike that, for *Pearson correlation* (r) absolute positioning is irrelevant.

While r formalizes the notion of "coherent" or "similar" behavior of models' layers, δ_F complements it with exact values similarity constraint (see Figure 2).

Frechet distance Given the simultaneous iterative step-by-step walkthrough from the start to the end points of both curves, one could freely vary the step size for every curve at every iteration. By the proper choice of step sizes during the walkthrough, one could guarantee that the optimal distance between curves' respective points will never be exceeded during the iteration process. That optimal distance is called Frechet distance and is formally calculated as follows: $\delta_F = \inf_{a,b}\{\max_t\{d(A_a(t), B_b(t))\}\}$, where t denotes iteration steps, a, b combinations correspond to various step size strategies, and A, B are the functions respective to the curves.

Pearson correlation coefficient Pearson correlation measures the strength of linear dependence of two samples: $r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y}$, where s_α is the standard deviation of sample α and $\overline{\alpha}$ is this sample mean.

3.4.2 Visualization

We provide the GUI (see Figure 3) to allow us to configure the similarity thresholds and explore the particular categories' results on a geospatial chart.

GUI allows setting off the δ_F and r absolute values thresholds and specifying particular languages and categories to be shown.

4 Evaluation Setup

To present the whole procedure of our probing framework working process, we decided to run the experiments only for two multilingual transformer encoder-based models: the 12-layer mBERT model (Devlin et al., 2018)[5] and the 24-layer XLM-RoBERTa model (Conneau et al., 2019)[6]. We used embeddings from "[CLS]" token for each text sample as it is widely accepted. As the classifier, while supporting LogReg and MLP, we choose Logistic Regression due to its higher *Selectivity* (Hewitt and Liang, 2019). The classifier was trained on 10 epochs using cross-entropy loss and *AdamW* (Loshchilov and Hutter, 2017) optimizer. A separate classifier was trained for each feature of all languages and each layer.

To eliminate the problem of different sizes of the datasets, we run the classifier five times and then take an average result to avoid the classifier bias. The results were evaluated by F_1 weighted score because of the unbalanced data for most probing tasks. From Universal Dependencies, using our SentEval converter, we obtained 1927 probing tasks for 104 languages. During the training, we noticed

[4] By probing curve we refer to the typical probing chart. Layers, or other probed parts of a model, and the respective results are visualized as a curve on a linear chart.

[5] https://huggingface.co/bert-base-multilingual-cased

[6] https://huggingface.co/xlm-roberta-large

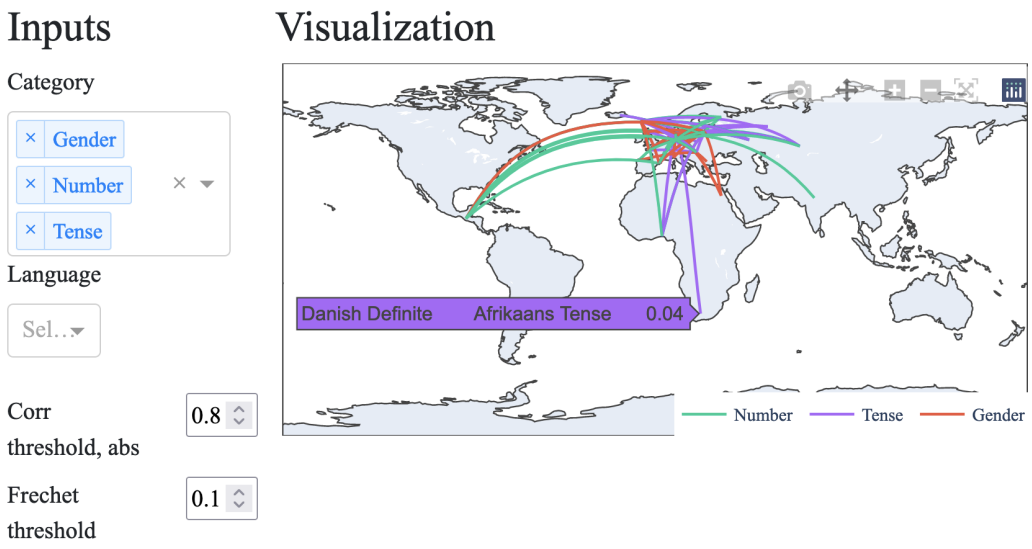

Figure 3: GUI screenshot: Similarity between languages learned by mBERT based on different probing tasks.

that some samples contain long sentences with token numbers of more than 512. We propose two options for handling it correctly: truncate sentences to 512 tokens or dispose of all of these sentences.

5 Results and Insights

5.1 General Results

We received a massive multilingual probing task bundle of 1927 tasks using all the converted data for 104 languages. It took us 10 hours to probe through all the files on one NVidia Tesla GPU V100.

We thus conducted the probing of the mBERT and XLM-R models to figure out the capabilities of the models, as follows:

1. to generalize linguistic information language-wise: grouping the average results a) by layers (Figure 5), b) by each feature in each language (Figure 4).
2. to generalize linguistic information feature-wise, grouping the average results by each layer and by each language (Appendix A.2).
3. to explore the results feature-wise: a) by searching for similarities in layer-wise feature representations) by exploring individual feature results grouped by language and layer (Appendices A.4, A.5), c) by creating the geospatial visualizations of the similar features.

The model evaluation results are presented in Figure 4: the figure clearly shows the sparseness with which all features are presented in each language. Basic features such as Number, PronType, and Tense are among the most frequent ones. The example of the geospatial visualizations of the similarly learned features is presented in Figure 3

5.2 mBERT and XLM-R Multilingual Abilities and Insights

Given the mBERT model as an example, as for the categories *Number* and *PronType*, which are the most common across languages, the best scores were achieved at the 3rd and 6th layers, respectively. In the case of all categories, mBERT showed the best result at the 5th layer. In Appendix A.4 and A.5 the heatmaps for all languages with these categories can be found. As for the average results by all categories, see Appendices A.2 and A.3. Figure 5 depicts average language scores for Number and PronType and among all categories.

As mentioned above, we evaluated two models, mBERT and XLM-R, on our data. On average, their performance is similar. However, the scores of XLM-R are slightly worse than the ones of mBERT. By the categories, mBERT shows the best performance in Hungarian, Chinese, Urdu, Welsh, and Slovak. The worst results were shown in Tupinambá and Akuntsu. XLM-R's top languages include the same language in a different order (Chinese has the best quality).

On Javanese and Akuntsu, XLM-R shows the worst quality. The models show the best quality on high-resource languages and have worse representations of under-resourced non-Indo-European languages.

Among the factors that can impact the models'

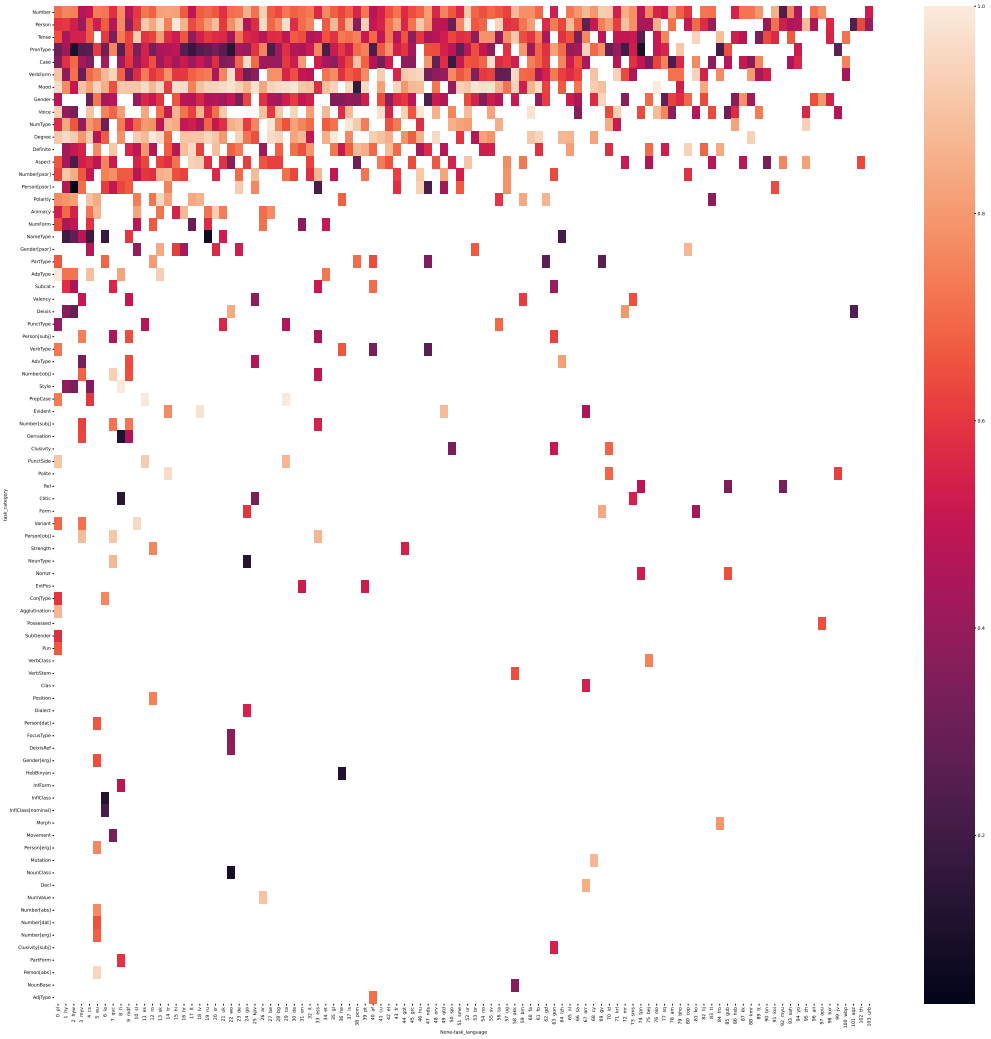

Figure 4: mBERT results grouped by languages and average feature probing score on all layers

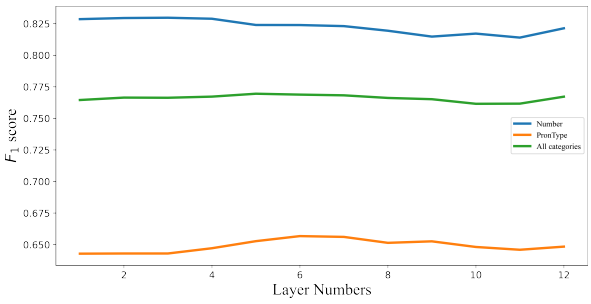

Figure 5: Distribution of scores by model's layers depending on the categories which are used across languages.

performance on different languages, the following might be the most essential: script, language genealogy, and typological features of languages.

To research the effect of script and language genealogy on the results, we run an ANOVA test since we have more than two families or scripts. The test reveals a strong correlation between a language family and the models' performance ($p = .0005$ for both XLM-R and mBERT).

We also run an ANOVA test to see if another significant difference in performance on languages with different scripts exists. The test shows that script did not impact the final performance ($p = .52$ for mBERT and $p = .39$ for XLM-R). The reported difference in the performance may be caused by the set of categories or the dataset size. Independently, other works (Pires et al., 2019; Wu and Dredze, 2019) claim that mBERT shows language neutrality regarding both a language and its script. Our

results support that level of performance does not depend on a script, as models show high results on languages with Arabic-based (Persian, Urdu) or Ge'ez scripts (Amharic).

Yet, the models might be biased towards Standard Average European languages (SAE) (Haspelmath, 2001), as it solves tasks on the categories found in SAE languages better than on the other language-specific categories. For example, the top-10 of the best recognised categories include *Person[abs], PunctSide, Person[obj], Agglutination* [7]*, Mutation, Degree, Decl, Mood* [8]*, Evident*[9], and *Polarity*[10]. Agglutination and Mutation are highly imbalanced towards one class, and other categories, except for Evident, are widespread in European languages.

On the other hand, top-10 worst categories are following: *NumValue, InflClass, NounClass, HebBinyan, Clitic, ExtPos, Derivation, NameType, InflClass[nominal], FocusType*. These categories are language-specific and are not found in SAE languages. Apart from that, if a category typical for SAE gets a different set of values, the model performs much worse. The model generally shows good results on *Case* and significantly worse results for Hungarian and Amharic with a different set of values for *Case*.

Chi et al. (2020) prove that mBERT has a joint subspace of universal syntactic relations. Since we cannot fully prove if mBERT has a joint subspace of morphological features because, as mentioned before, some morphological features are not universal. However, we can see if there is any correlation between categories across all languages based on how different layers learn these features, according to Frechet distance. Most categories do not have a correlating category. However, there are several compelling cases to mention. mBERT generally learns Evident and Mood similarly, and in some languages, such as Bulgarian, Evident is regarded as a value of *Mood*. Other than that, Definite and Number have a little distance, which might be expressed with one morpheme, as in Scandinavian languages. The same is valid for Number and Person categories that are learned similarly by mBERT.

There is not enough evidence to claim that mBERT has a joint subspace for morphological features because categories have different sets of values, and mBERT performs better on SAE categories. Yet, it shows some generalization abilities on the similarity of morphology across languages.

6 Future Work

As part of future work, we plan to include syntactic markup in research and an interface with a CQL-like[11] query language for authors. The ability to set conditions on a subcorpus of examples will give the researchers the freedom to create custom and linguistically motivated probing tasks while the rest of the experiment parameters will be fixed. One can imagine, e.g., the probing of the model on the `[tag="NP"]` query, exploring the results specifically on the noun phrases. We believe that, in this respect, the tasks of probing and interpreting the results of the model become close to the tasks of corpus linguistics and searching through corpora for statistical testing of hypotheses.

7 Conclusion

The typological variety of linguistic features composes the general nature of language. To address the lower-abstract parts of this nature, we introduced the Universal Probing framework, which allows the researchers to run and aggregate massive amounts of probing experiments in a fixed and reproducible setup.

The current framework version includes an experimental setup from 104 languages and 80 grammatical features. The framework can be used for language model interpretation with various architectures, and the results can also be easily incorporated into the model cards checklist. It can be used in more language-wise transfer learning and typological studies with multilingual models.

We hope that the community will use our work in order to interpret, evaluate and compare the language models, leading to better and more explainable NLP. The framework and all the data are open-source under Apache 2.0 license `https://github.com/AIRI-Institute/Probing_framework`.

[7] Only used for Polish past participles

[8] Mood express modality, such as indicative, imperative, conditional

[9] Evidence is the morphological marking of a speaker's source of information (Aikhenvald, 2006)

[10] Polarity shows if words can be used only in negative or positive contexts

[11] `https://www.sketchengine.eu/documentation/corpus-querying/`

8 Limitations

By now, it is also worth mentioning the UD data dependencies of the framework. The problems in the UD data, such as annotation errors, formatting errors, and version instability, could potentially affect the resulting probing framework. As described in Section 3.2, we have eliminated obviously problematic fragments; however, more deeply incorporated inaccuracies may drag on, surviving conversion to the SentEval format. Some inconsistencies in the accepted annotation format affected the quality of model embeddings: such categories as PunctSide (Catalan, Finnish, Icelandic, Polish, Spanish), NameType (Armenian, Classical Chinese, Czech, Erzya, etc.) are rare and have a very different distribution from language to language and are expected to be at the bottom of the list. The categories accepted in the UD for one specific language are also poorly solved: Agglutination (Polish), Mutation (Welsh), HebBinyan (Hebrew), NounClass (Wolof), NumValue (Arabic, Czech).

We use the latest available UD release (version 2.10)[12]. As stated on the project's website, the next release (v2.11) is scheduled for November 15, 2022, so data curation and updates will be necessary to incorporate the newer and better UD annotation into the framework.

The proposed framework allows for different probing methods to be used similarly, including the widely criticized ones (Belinkov, 2022). Researchers relying on the presented framework should carefully pick the proper methods in their probing studies. For example, we've introduced the control task (Hewitt and Liang, 2019) consisting of averaging the probing performance across several probing experiments. This reduced the possibility of a probing task erroneously receiving a high score due to the small size of the testing data.

9 Ethical Considerations

9.1 Possible Misuse

The framework's usage implies working concerning standard practices during model pre-training, such as controlling that the test data (e.g., UD corpora) are excluded from the training corpus. Using UD data during pre-training or fine-tuning the model can lead to indicative and biased results of model interpretation.

9.2 Data-specific Problems

9.2.1 Dataset Characteristics

The dataset covers the languages described in Section A.1. The probing dataset statistics are also presented in Section A.1.

9.2.2 Generalization

The UD data can be considered mostly validated, as it involves multiple institutions to develop and test the annotation standards, as well as the corpus data itself. However, besides data quality, usage of the data should address such characteristics as quantity: that is why we have automatically excluded the UD categories having only one value within a category in all available languages. For all other categories, the data for the classification task were not limited in any way; the train/val/test data division was preserved.

Potentially, other data-dependent problems (see also the resulting data dependencies in Section 8) could be:

- genre bias in specific languages;
- personal style/resource bias in specific languages;
- collocation of the specific features: some features can possibly occur within the same contexts (sentences), which makes the solution of the classification problem within the probing setup noisy for the tested language model.

Nevertheless, we consider the UD corpora to be sufficiently reliable and the most complete of the available data for a detailed low-level multilingual probing study of the models.

9.2.3 Data Quality

In addition to the above, we draw attention to the question of the language representation problems in the UD. According to the Ethnologue database [13], there are more than 4000 languages with developed writing systems, while only 172 of them are presented in the UD in general, and even less (104) were qualified to be included in the framework format.

As we understand that the presented language set is not typologically sampled, we proceeded from the criterion of completeness, not balance. If necessary, we encourage willing researchers to sample their subsamples from our data to follow typological sampling.

[12] Version 2.10 treebanks are available at http://hdl.handle.net/11234/1-4758. 228 treebanks, 130 languages, released May 15, 2022.

[13] https://www.ethnologue.com/enterprise-faq/how-many-languages-world-are-unwritten-0

References

Alexandra Y Aikhenvald. 2006. Evidentiality. oxford: Oxford university press, 2004.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Ethan A Chi, John Hewitt, and Christopher D Manning. 2020. Finding universal grammatical relations in multilingual bert. *arXiv preprint arXiv:2005.04511*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. *ArXiv*, abs/1803.05449.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Fahim Dalvi, Avery Nortonsmith, D. Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, and James Glass. 2019. Neurox: A toolkit for analyzing individual neurons in neural networks. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Mohsen Fayyaz, Ehsan Aghazadeh, Ali Modarressi, Hosein Mohebbi, and Mohammad Taher Pilehvar. 2021. Not all models localize linguistic knowledge in the same place: A layer-wise probing on BERToids' representations. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 375–388, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Martin Haspelmath. 2001. The european linguistic area: Standard average european. In Wulf Oesterreicher Martin Haspelmath and Wolfgang Raible, editors, *Language Typology and Language Universals, Handbücher zur Sprach- und Kommunikationswissenschaft*, pages 1492–1510. Mouton de Gruyter, Berlin.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. *ArXiv*, abs/1909.03368.

Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. 2020. exbert: A visual analysis tool to explore learned representations in transformer models. In *ACL*.

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.

Fajri Koto, Jey Han Lau, and Tim Baldwin. 2021. Discourse probing of pretrained language models. In *NAACL*.

Murathan Kurfalı and Robert Östling. 2021. Probing multilingual language models for discourse. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 8–19, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Andreas Madsen, Siva Reddy, and A. P. Sarath Chandar. 2021. Post-hoc interpretability for neural nlp: A survey. *ArXiv*, abs/2108.04840.

Paola Merlo. 2019. Probing word and sentence embeddings for long-distance dependencies effects in French and English. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 158–172, Florence, Italy. Association for Computational Linguistics.

Vladislav Mikhailov, Ekaterina Taktasheva, Elina Sigdel, and Ekaterina Artemova. 2021. RuSentEval: Linguistic source, encoder force! In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 43–65, Kiyv, Ukraine. Association for Computational Linguistics.

Johanna Nichols. 2007. What, if anything, is typology?

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Vinit Ravishankar, Lilja Øvrelid, and Erik Velldal. 2019a. Probing multilingual sentence representations with x-probe. In *RepL4NLP@ACL*.

Vinit Ravishankar, Lilja Øvrelid, and Erik Velldal. 2019b. Probing multilingual sentence representations with X-probe. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 156–168, Florence, Italy. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Gözde Gül Sahin, Clara Vania, Ilia Kuznetsov, and Iryna Gurevych. 2020. Linspector: Multilingual probing tasks for word representations. *Computational Linguistics*, 46:335–385.

Maria Tikhonova, Vladislav Mikhailov, Dina Pisarevskaya, Valentin Malykh, and Tatiana Shavrina. 2022. Ad astra or astray: Exploring linguistic knowledge of multilingual bert through nli task. *Natural Language Engineering*, page 1–30.

Ekaterina Voloshina, Oleg Serikov, and Tatiana Shavrina. 2022. Is neural language acquisition similar to natural?a chronological probing study.

Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. 2019. Allennlp interpret: A framework for explaining predictions of nlp models. *ArXiv*, abs/1909.09251.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. *arXiv preprint arXiv:1904.09077*.

Gözde Gül Şahin, Clara Vania, Ilia Kuznetsov, and Iryna Gurevych. 2019. Linspector: Multilingual probing tasks for word representations.

A Appendix

A.1 Languages information and statistic

After the processing of Universal Dependencies, 104 languages left. Here is a table with the languages, their families, and the number of examples that we used in the experiments.

Language	Family	Examples	Language	Family	Examples
Afrikaans	Indo-European	19646	Komi Zyrian	Uralic	5682
Akkadian	Afro-Asiatic	15037	Korean	Koreanic	3314
Akuntsu	Tupian	79	Kurmanji	Indo-European	9134
Albanian	Indo-European	380	Latin	Indo-European	1048162
Amharic	Afro-Asiatic	7166	Latvian	Indo-European	393694
Ancient Greek	Indo-European	566076	Ligurian	Indo-European	2304
Apurina	Arawakan	176	Lithuanian	Indo-European	81462
Arabic	Afro-Asiatic	484604	Livvi	Uralic	835
Armenian	Indo-European	37117	Low Saxon	Indo-European	623
Assyrian	Afro-Asiatic	152	Manx	Indo-European	13536
Bambara	Mande	4829	Marathi	Indo-European	6340
Basque	-	191646	Mbya Guarani	Tupian	5503
Beja	Afro-Asiatic	347	Moksha	Uralic	3133
Belarusian	Indo-European	445666	Munduruku	Tupian	164
Bengali	Indo-European	156	Naija	Atlantic-Congo	42928
Bhojpuri	Indo-European	1525	North Sami	Uralic	21639
Breton	Indo-European	5206	Norwegian	Indo-European	631651
Bulgarian	Indo-European	238822	Old Church Slavonic	Indo-European	118508
Buryat	Mongolic-Khitan	6832	Old East Slavic	Indo-European	153393
Catalan	Indo-European	305522	Old French	Indo-European	45115
Chinese	Sino-Tibetan	18865	Persian	Indo-European	183678
Classical Chinese	Sino-Tibetan	93864	Polish	Indo-European	860418
Coptic	Afro-Asiatic	22150	Portuguese	Indo-European	197481
Croatian	Indo-European	193156	Romanian	Indo-European	543203
Czech	Indo-European	831540	Russian	Indo-European	270189
Danish	Indo-European	104906	Sanskrit	Indo-European	25885
Dutch	Indo-European	241808	Scottish Gaelic	Indo-European	27907
English	Indo-European	414215	Serbian	Indo-European	94856
Erzya	Uralic	17458	Skolt Sami	Uralic	989
Estonian	Uralic	557773	Slovak	Indo-European	218032
Faroese	Indo-European	21133	Slovenian	Indo-European	286196
Finnish	Uralic	624845	Spanish	Indo-European	660046
French	Indo-European	686410	Swedish	Indo-European	213496
Galician	Indo-European	15878	Tagalog	Austronesian	380
German	Indo-European	311259	Tamil	Dravidian	12602
Gothic	Indo-European	99064	Tatar	Turkic	250
Greek	Indo-European	49364	Thai	Tai-Kadai	612
Guajajara	Tupian	409	Tupinamba	Tupian	593
Hebrew	Afro-Asiatic	112866	Turkish	Turkic	746291
Hindi	Indo-European	321197	Turkish German	Indo-European	21160
Hungarian	Uralic	41102	Ukrainian	Indo-European	139882
Icelandic	Indo-European	503790	Upper Sorbian	Indo-European	5241
Indonesian	Austronesian	47426	Urdu	Indo-European	96902
Irish	Indo-European	93264	Uyghur	Turkic	40174
Italian	Indo-European	395470	Warlpiri	Pama-Nyungan	128
Javanese	Austronesian	290	Welsh	Indo-European	17026
Kaapor	Tupian	99	Western Armenian	Indo-European	71303
Karelian	Uralic	1475	Wolof	Atlantic-Congo	21518
Karo	Tupian	1845	Xibe	Tungusic	4226
Kazakh	Turkic	15082	Yakut	Turkic	213
Kiche	Indo-European	11534	Yoruba	Atlantic-Congo	1151
Komi Permyak	Uralic	325	Yupik	Eskimo-Aleut	2281

A.2 mBERT's layers F_1 scores for all languages and categories.

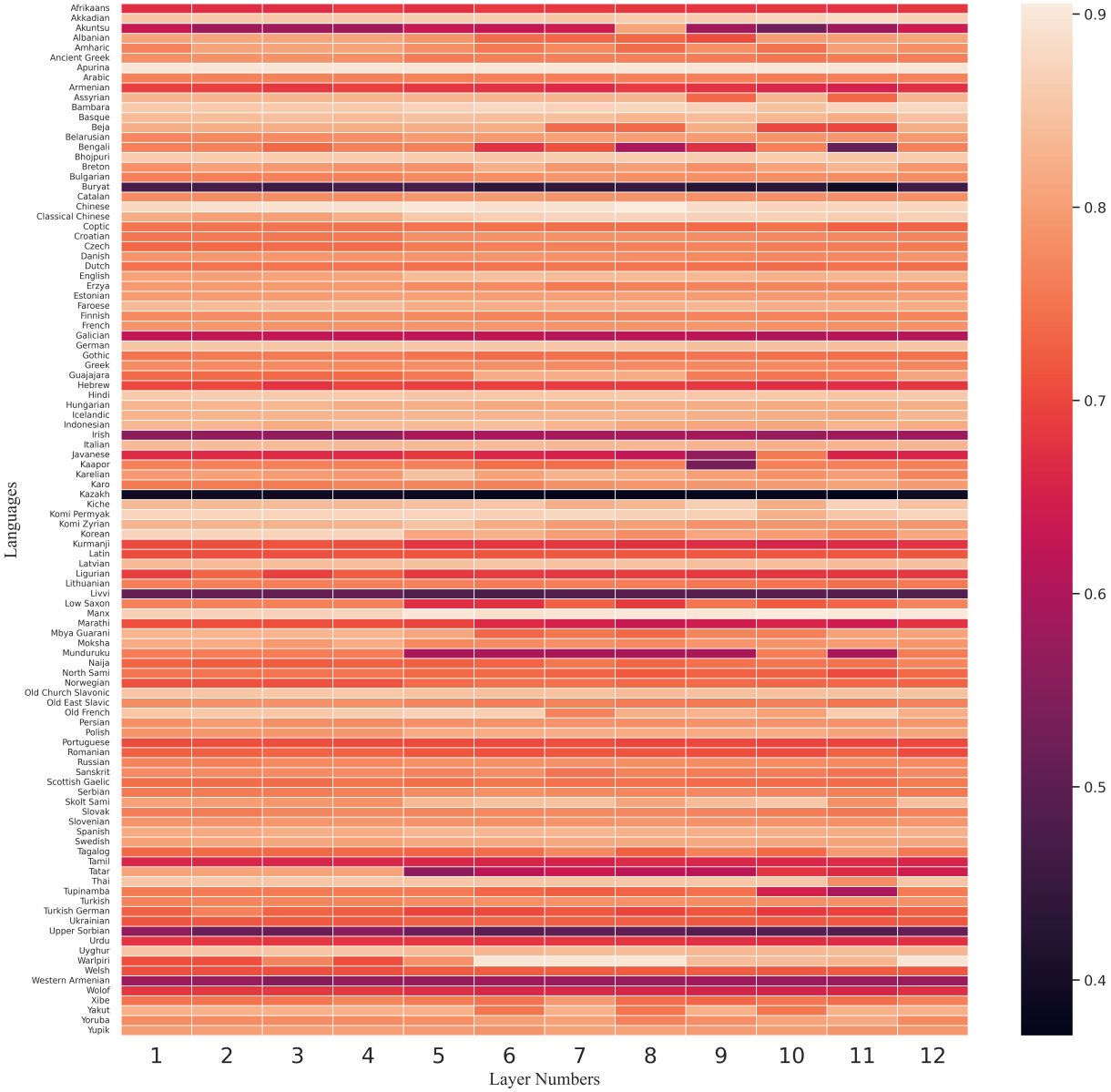

Figure 6: Distribution of average model scores by layers for languages measured on all categories.

A.3 XLM-R results grouped by languages and average feature probing score on all layers

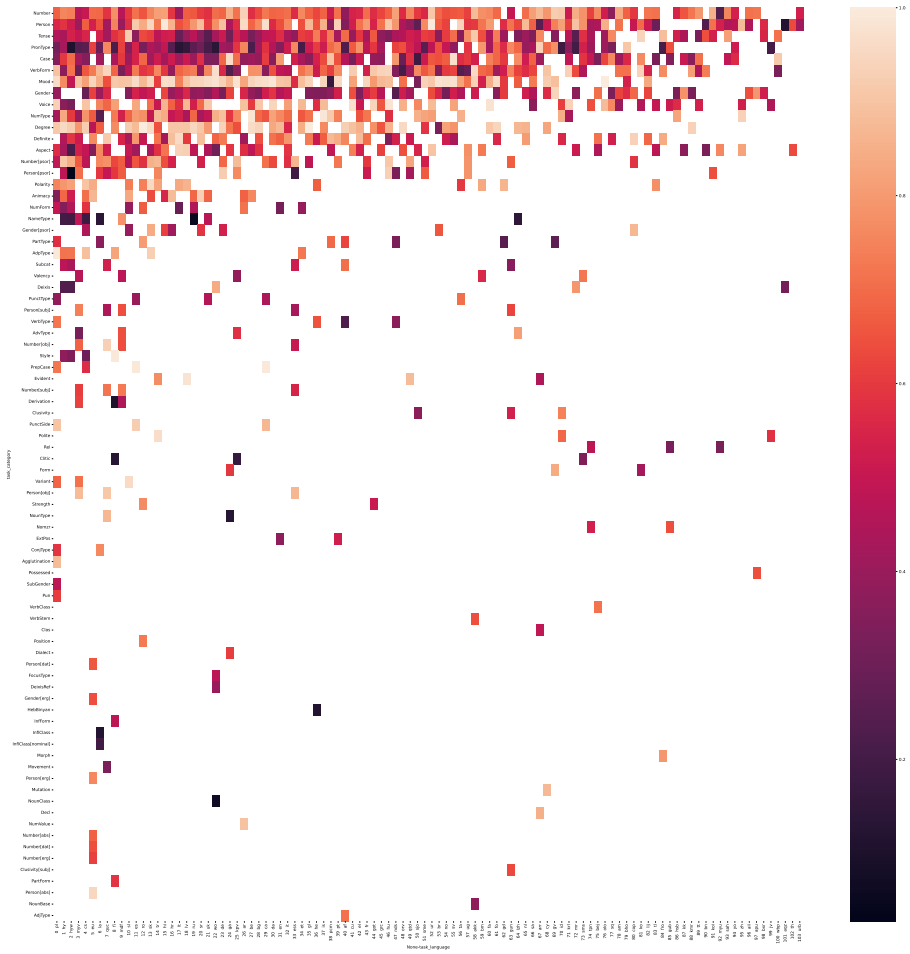

Figure 7: XLM-R results grouped by languages and average feature probing score on all layers

A.4 Model's layers F_1 scores for all languages on category "Number".

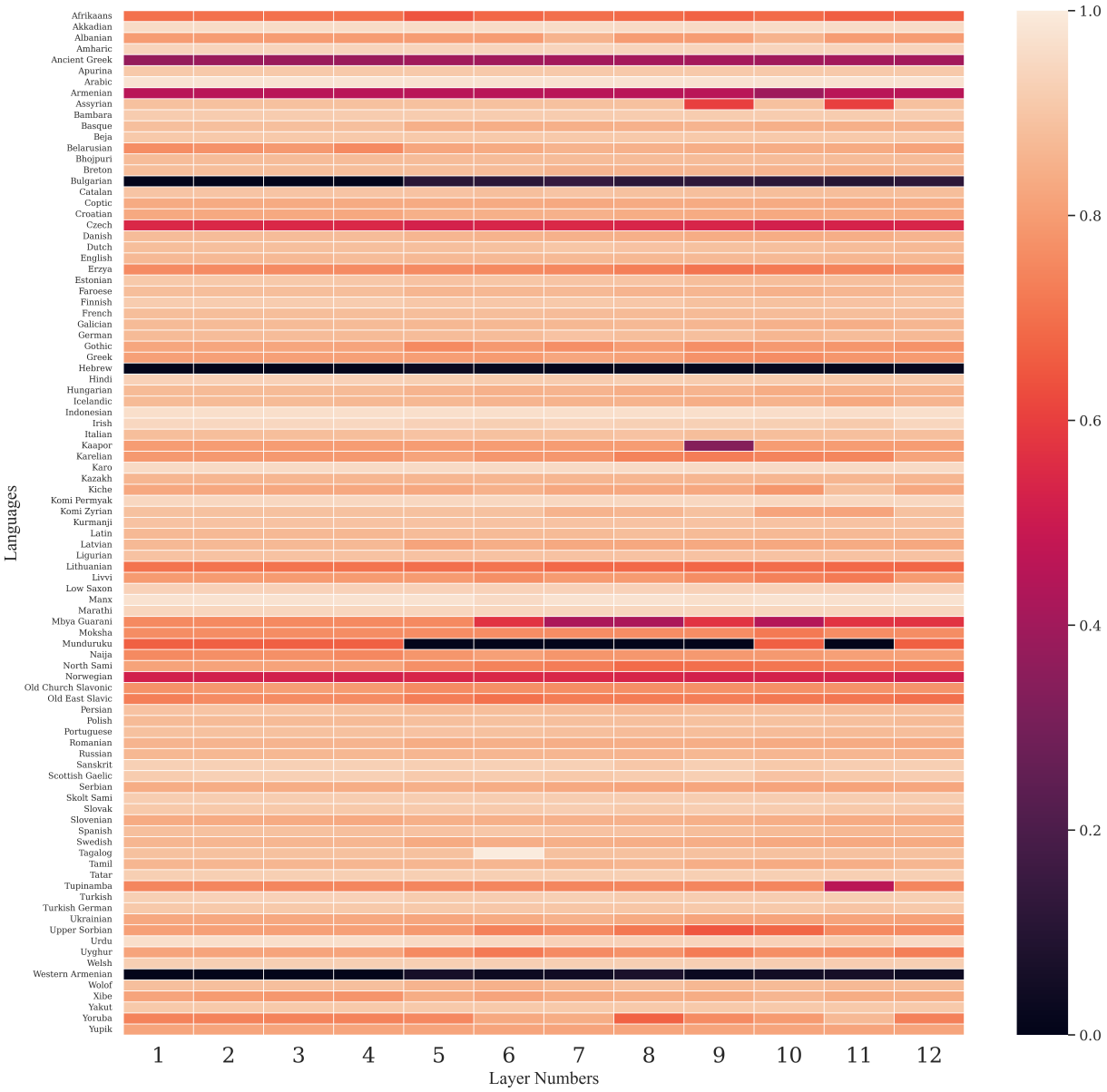

Figure 8: Distribution of model scores by layers for languages measured on category *Number*.

A.5 Model's layers F_1 scores for all languages on category "PronType".

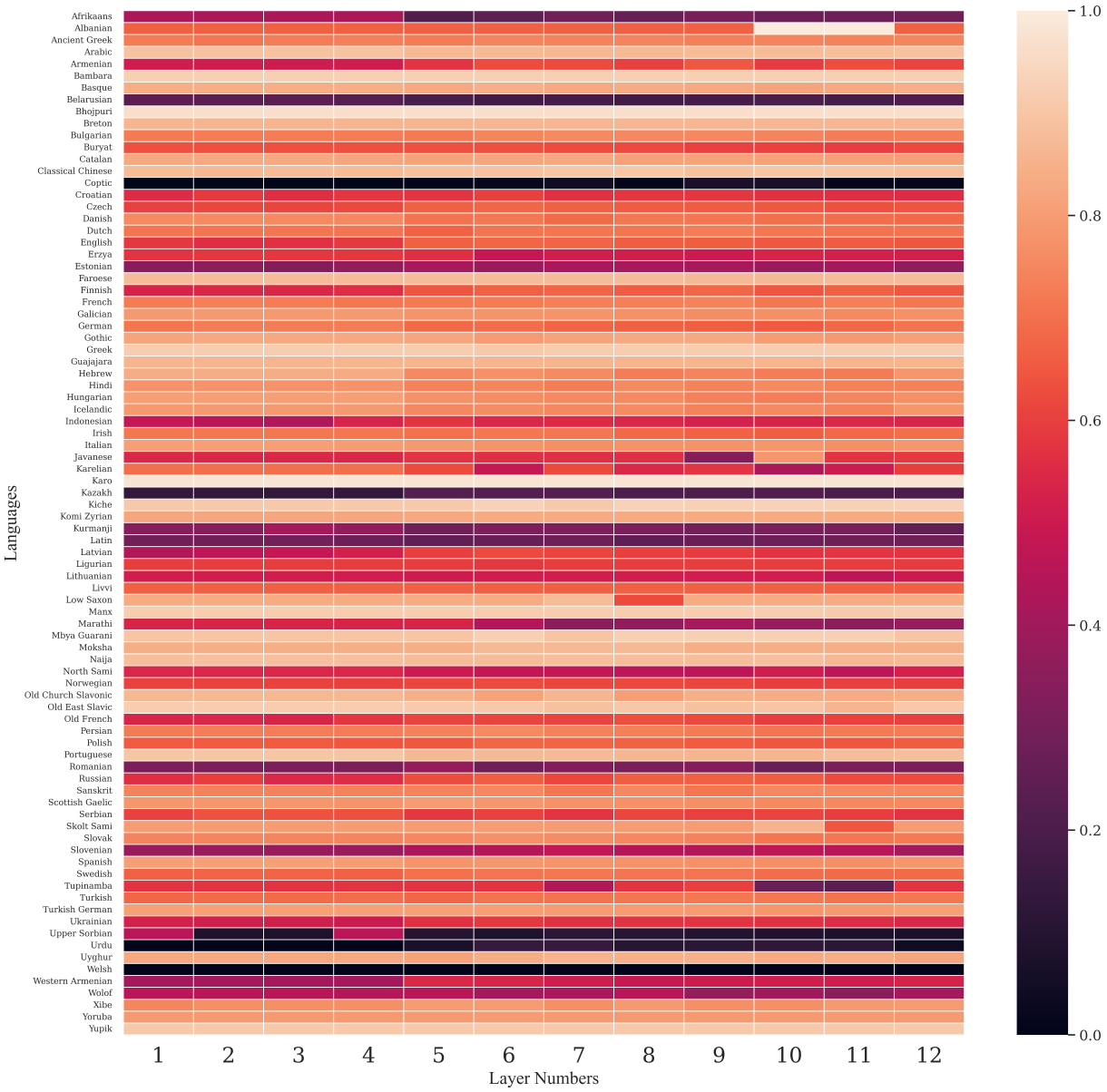

Figure 9: Distribution of model scores by layers for languages measured on category *PronType*.