A multilabel approach to morphosyntactic probing

Naomi Tachikawa Shapiro **Amandalynne Paullada** **Shane Steinert-Threlkeld**
Department of Linguistics, University of Washington, Seattle, WA, USA
{tsnaomi,paullada,shanest}@uw.edu

Abstract

We propose using a *multilabel* probing task to assess the morphosyntactic representations of multilingual word embeddings. This tweak on canonical probing makes it easy to explore morphosyntactic representations, both holistically and at the level of individual features (e.g., gender, number, case), and leads more naturally to the study of how language models handle co-occurring features (e.g., agreement phenomena). We demonstrate this task with multilingual BERT (Devlin et al., 2018), training probes for seven typologically diverse languages: Afrikaans, Croatian, Finnish, Hebrew, Korean, Spanish, and Turkish. Through this simple but robust paradigm, we verify that multilingual BERT renders many morphosyntactic features simultaneously extractable. We further evaluate the probes on six held-out languages: Arabic, Chinese, Marathi, Slovenian, Tagalog, and Yorùbá. This zero-shot style of probing has the added benefit of revealing which cross-linguistic properties a language model recognizes as being shared by multiple languages.

1 Introduction

Morphologically rich languages present unique challenges to natural language processing. These languages typically exhibit complex agreement patterns and their high diversity of inflected forms can lead to sparse examples of vocabulary words in training data, even in large corpora (Blevins and Zettlemoyer, 2019; Gerz et al., 2018). It is therefore worthwhile to explore how neural language models (LMs), which serve as the foundation of many state-of-the-art systems, handle the morphological complexity of diverse languages.

Morphosyntactic features of natural languages bear meaningful information that is useful for downstream tasks, such as machine translation, question answering, and language generation. Adding morphological supervision through multitask training regimes (Blevins and Zettlemoyer, 2019) or morphologically-informed tokenization (Klein and Tsarfaty, 2020; Park et al., 2020) can improve the quality of multilingual language models. Nonetheless, recent work has shown that LMs trained without explicit morphological supervision can still produce useful representations that capture morphosyntactic phenomena (e.g., Bacon and Regier, 2019; Pires et al., 2019; Dufter and Schütze, 2020).

To further these investigations, we propose using a multilabel probing task to assess the morphosyntactic representations of multilingual word embeddings. This work is premised on the intuition that, if a simple model (a "probe") can easily extract linguistic properties from embeddings, this indicates that the LM has learned to encode those features in some fashion (Conneau et al., 2018; Hupkes et al., 2018; Liu et al., 2019). We show how a multilabel paradigm can shed light on the morphosyntactic representations of LMs, both holistically and at the level of individual features.

Our contributions are threefold: First, we introduce an efficient probing paradigm for analyzing multiple morphosyntactic features, which we demonstrate with multilingual BERT (Devlin et al., 2018) and seven typologically diverse languages: Afrikaans, Croatian, Finnish, Hebrew, Korean, Spanish, and Turkish. Second, we evaluate the probes on six "held-out" languages—Arabic, Chinese, Marathi, Slovenian, Tagalog, and Yorùbá—showing how this paradigm can be used in a zero-shot manner to illuminate the properties that multilingual BERT represents similarly cross-linguistically. Third, we release our code and multilabel probe predictions to guide future probing efforts and to serve as the foundation for future in-depth feature-level analyses.[1]

This paper is structured as follows: Section 2

[1] https://github.com/tsnaomi/morph-bert

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4486–4524
November 7–11, 2021. ©2021 Association for Computational Linguistics

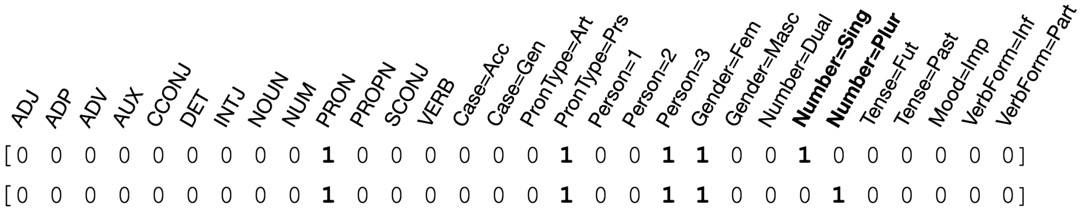


Figure 1: Hypothetical multi-hot representations of the Hebrew *3.sing.fem* pronoun היא *hi* (top) and *3.plur.fem* pronoun הן *hen* (bottom). The two vectors differ only with respect to the two cells indicating number.

reviews related work, motivating §3, which introduces multilabel morphosyntactic probing. Section 4 then outlines the data and models we use to probe multilingual BERT. In §5, we demonstrate the probing paradigm in a set of *monolingual* experiments, training and evaluating separate probes for the seven languages, and provide an example feature-level analysis of Hebrew determiners. In §6, we delve into whether *multilingual* probes yield comparable insights to the monolingual probes. Then, in a set of *crosslingual* experiments, §7 evaluates how the monolingual and multilingual probes handle the six held-out languages. Finally, §8 discusses our findings and concludes the paper.

2 Related work

Numerous studies in recent years have sought to study the linguistic properties captured by neural language models (e.g., Conneau et al., 2018; Gulordava et al., 2018; Hupkes et al., 2018; Marvin and Linzen, 2018; Zhang and Bowman, 2018; Bacon and Regier, 2019; Futrell and Levy, 2019; Hewitt and Manning, 2019; Jawahar et al., 2019; Liu et al., 2019; Tenney et al., 2019; Chi et al., 2020).

In the morphology domain, the LINSPECTOR suite by Şahin et al. (2020) probes 24 languages via 15 linguistic tasks, including multiple tasks to identify morphological features. In a similar vein, Edmiston (2020) uses several morphological prediction tasks to inspect embeddings from five monolingual Transformer-based language models, focusing exclusively on Indo-European languages. The probing paradigm proposed in this paper builds on these works, but consolidates morphosyntactic feature prediction under a single task that leads more naturally to the study of feature co-occurrence.

Recent probing work has also sought to curtail how much probes memorize about linguistic tasks to ensure that they *reflect* information available in their input embeddings—probes should be extractive rather than learnèd themselves. Efforts to minimize memorization have included reducing the training data to probes (Zhang and Bowman, 2018) and limiting probe complexity, such as through dropout (e.g., Belinkov et al., 2017a,b; Şahin et al., 2020) and the use of simpler architectures (e.g., a linear layer instead of a multilayer perceptron, as in Alain and Bengio 2018 and Liu et al. 2019).

To guide the design and interpretation of probes, Hewitt and Liang (2019) propose supplementing diagnostic tasks with *control tasks*, where a probe is trained to predict random outputs within the same output space as the diagnostic task, given the same embeddings. If the probe performs well on the control task, they caution that it has the capacity to memorize the linguistic features under consideration; conversely, if the probe does well on the diagnostic task but poorly on the control task, then it is a reliable diagnostic of linguistic representations in the embeddings (though see Pimentel et al. 2020a,b for interesting discussions). Hewitt and Liang operationalize this comparison as *selectivity*, the difference in performance between the diagnostic and control tasks. The greater the selectivity, the more the probe “expresses” the information encoded in its input. In this paper, we design a control task to complement multilabel probing.

3 Multilabel morphosyntactic probing

We propose using multilabel morphosyntactic tagging to assess the morphosyntactic representations of neural LMs. In this diagnostic task, we hold contextualized word embeddings constant, then train linear classifiers on top of them (cf. Liu et al., 2019; Hupkes et al., 2018) to perform morphosyntactic tagging. In its objective, morphosyntactic tagging resembles the second SIGMORPHON 2019 shared task, which called for labeling words in a sentence with their morphosyntactic descriptions (McCarthy et al., 2019).

It is easy to imagine doing morphosyntactic tagging in a traditional multiclass fashion, where we train separate probes to identify different features, such as part of speech (POS), gender, or number (cf. Şahin et al., 2020; Edmiston, 2020). However, this style of probing is more likely to prompt narrow analyses that consider morphological properties in isolation. Alternatively, we could train a single probe to extract complex labels like `def.sing.masc.noun` and `3rd.plur.masc.past.verb`. Thus, each word would have a single correct label and a final softmax layer would output the probability of each class being the correct one. However, a drawback to this approach is that, depending on the number of properties we would like to identify, this can result in a combinatoric nightmare 👻, with few training examples per class.

To overcome these limitations, we frame morphosyntactic tagging as a word-level multilabel task, allowing for a token to receive multiple feature labels (e.g., both `Person=1` and `Number=Sing`) that are multi-hot encoded. Such a paradigm allows us to encode features with multiple or ambiguous values (e.g., `Gender=Fem,Masc`; a.k.a. multi-valued features) and enables a closer inspection of learnt agreement and feature co-occurrence patterns. Figure 1 illustrates hypothetical gold vectors for two Hebrew pronouns that differ only in number.

3.1 Notation and nomenclature

We define a feature label as the conjunction of a linguistic feature (e.g., *number*) and a possible realized value of that feature (e.g., *singular*). Multiple feature labels can correspond to the same feature (e.g., `Number=Sing` and `Number=Plur`).[2] We define F as the set of feature labels $\{f_1, \dots, f_{|F|}\}$ that we use to identify morphosyntactic properties from word embeddings.

Assuming a vocabulary of word types V, let $\mathbf{s} = s_1 \dots s_{|\mathbf{s}|}$ denote a specific sentence and r_i denote the contextualized representation of each token s_i, such that $s_i \in V$. The inputs to the probe are therefore the embeddings $r_i \in \mathbb{R}^d$. In the multilabel morphosyntactic tagging task, we define the target output of each embedding r_i as a multi-hot encoded vector $\mathbf{y}^i = y^i_1 \dots y^i_{|F|}$, where F is the aforementioned set of feature labels. We encode y^i_j as 1 if the feature label $f_j \in F$ describes the token s_i and 0 otherwise.

3.2 Multilabel evaluation

The multilabel paradigm lends itself well to analyzing features both holistically and at a granular level. We can analyze individual features by calculating precision, recall, and F_1 for each feature label f separately. Furthermore, we can glean the overall or *micro-averaged* performance of a probe by first tallying the true positives (TP), false positives (FP), and false negatives (FN) across the features, before calculating precision, recall, and F_1.

4 Experimental setup

We demonstrate multilabel morphosyntactic probing with multilingual BERT (henceforth, mBERT; Devlin et al., 2018), using morphologically annotated corpora from Universal Dependencies (UD; Nivre et al., 2016, 2020).[3]

4.1 Data

In a set of monolingual experiments, we trained separate probes to predict morphosyntactic features from corpora for seven languages of varying morphological complexity: Afrikaans (AfriBooms; cf. Dirix et al., 2017), Croatian (SET; cf. Agić and Ljubešić, 2015), Finnish (TDT; cf. Haverinen et al., 2014; Pyysalo et al., 2015), Hebrew (HTB; cf. Tsarfaty, 2013; McDonald et al., 2013; Sadde et al., 2018), Korean (PUD; cf. Zeman et al., 2017), Spanish (AnCora; cf. Alonso and Zeman, 2016), and Turkish (IMST; cf. Sulubacak et al., 2016; Tyers et al., 2017; Türk et al., 2019). With the exception of the Korean data, all of the corpora came pre-split into training, validation, and test sets. We performed an 80-10-10 split on the 1,000-sentence Korean PUD corpus. To throttle the probes' training data (cf. Zhang and Bowman, 2018), we reduced the other training sets to 800 sentences as well.

Next, in a set of multilingual experiments, we trained probes on a shuffled combination of the training sentences from the monolingual probes. However, we excluded the Korean dataset from this analysis, due to the lack of documentation on its construction. Finally, in a set of crosslingual transfer experiments, we evaluated the monolingual and multilingual probes on six held-out languages: Arabic (PADT; cf. Smrž et al., 2002, 2008; Hajič et al., 2009), Chinese (PUD; cf. Zeman et al., 2017), Marathi (UFAL; cf. Ravishankar, 2017), Slovenian

[2] We drop `POS=` from part-of-speech labels, conforming to UPOS notation (e.g., `NOUN` instead of `POS=NOUN`).

[3] https://universaldependencies.org/introduction.html

(SST; cf. Dobrovoljc and Nivre, 2016), Tagalog (TRG), and Yorùbá (YTB; cf. Ishola and Zeman, 2020). To clarify, mBERT *was* pre-trained on these languages; we consider them "held-out" in that we never train probes to extract linguistic properties from these corpora (i.e., the experiments are zero-shot). The datasets for the monolingual, multilingual, and crosslingual experiments are summarized in Appendix A.

All of the probes were trained to extract multiple features, such as POS, number, gender, case, and tense, as well as language-specific features, such as Finnish infinitive forms. (It is due to the inclusion of parts of speech that we refer to the task as "morpho*syntactic* tagging".) Since the languages vary in their linguistic properties, we used different label sets for each language and a semi-aggregated set for the multilingual probes. Across our experiments, we extracted 166 different feature labels in total, as listed in Appendix B.

The UD corpora include decompositions of multiword tokens and separate annotations for their respective components. To keep the input to the probes faithful to naturalistic text, we embed the multiword tokens themselves, but aggregate the feature labels from their components (e.g., the Hebrew multiword token הספר *hasefer* 'the book' is marked as both a determiner and noun).

4.2 Models and training

For our experiments, we instantiated a "BERT-Base, Multilingual Cased" model using HuggingFace's *Transformers* library (Wolf et al., 2019). This BERT variant contains 110M parameters across 12 Transformer layers, each with 12 attention heads and a hidden size of 768. The model was pre-trained on Wikipedia dumps from 104 languages. The authors over-sampled the smaller Wikipedia corpora to create a more cross-linguistic vocabulary, consisting of 100K wordpieces.

We froze mBERT and trained linear classifiers on top of embeddings produced by mBERT's initial embedding layer and its successive Transformer layers (cf. Liu et al., 2019; Hupkes et al., 2018). Preliminary experiments showed that the even-numbered layers (mBERT-0, mBERT-2, mBERT-4, etc.) faithfully captured the layer-by-layer trends across mBERT, so we opted to cut down on computation by focusing exclusively on these layers. The classifiers used sigmoid activation and were trained with mean binary cross-entropy loss to perform the multilabel tagging task. We trained each classifier for 50 epochs, selecting the model from the epoch that achieved the best validation loss. Courtesy of PyTorch (Paszke et al., 2019), the classifiers were optimized using Adam (learning rate = 0.001, β_1=0.9, β_2=0.999, ϵ=1e-08; Kingma and Ba, 2015). No dropout was used.

We performed word-level predictions of morphosyntactic properties by first summing over the word-piece embeddings for each word, then caching these representations prior to training the probes. See Appendix C for more details.

4.3 Vying for control

Following Hewitt and Liang (2019), we constructed a control task to complement the multilabel tagging task, whereby each word type in the task vocabulary was assigned a multi-hot output vector that was randomly generated according to the true distribution of feature labels in the training data. Deviating from Hewitt and Liang's notation, we generated a control output vector $\mathbf{c^i}$ for each word type $v_i \in V$, such that $\mathbf{c}^i = c_1^i \dots c_{|F|}^i$, where c_j^i was sampled from the true distribution of feature f_j in the training data. For instance, if f_j was a feature of 4% of the tokens in the training set, then c_j^i has a 0.04 probability of being 1 for any word type v_i (or, conversely, a 0.96 probability of being 0).

5 Monolingual experiments

In a set of monolingual experiments, we trained and evaluated individual diagnostic probes on Afrikaans, Croatian, Finnish, Hebrew, Korean, Spanish, and Turkish, given representations from the even-numbered mBERT layers. Their micro-averaged F_1 scores are conveyed in Figure 2, along with their results on the analogous control tasks.

5.1 Monolingual performance at a glance

The micro-averaged F_1 scores confirm that mBERT renders many morphosyntactic properties easily extractable, with the best performing probes for each language achieving scores between 0.83 and 0.97. We find that mBERT-6 scored the highest across the languages. This is consistent with prior work that has shown English BERT's interior layers to perform best on similar linguistic tasks (Liu et al., 2019; Tenney et al., 2019). Once mBERT has encoded morphologically relevant information, it seems that probe performance steadily declines as the topmost layers gear up for cloze predictions.

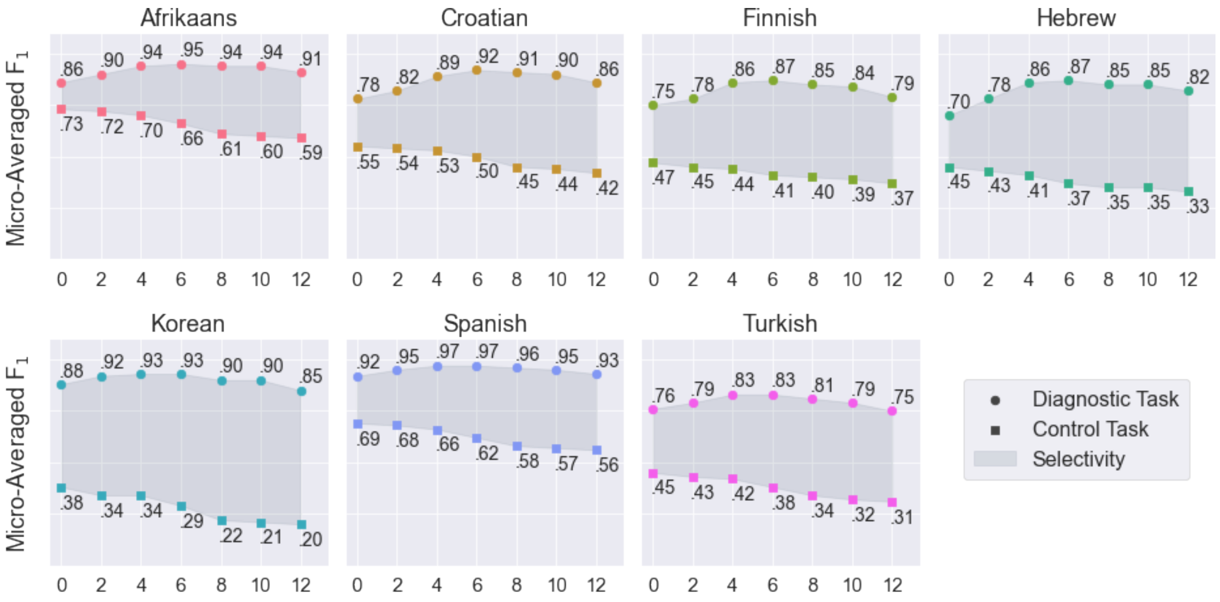


Figure 2: Micro-averaged F_1 results from the monolingual probes on the diagnostic and control tasks. The x-axes indicate the mBERT layer.

Notably, the Afrikaans and Spanish probes performed the best and the Turkish probes the worst. It is tempting to conclude that 'mBERT knows Afrikaans and Spanish better than Turkish'. However, we should refrain from comparing global probe performance across languages, as each language differed in the sets of features that were extracted. Furthermore, although each of the probes were trained on 800 sentences, they were ultimately trained on varying numbers of tokens. It may be that the Afrikaans and Spanish probes performed the best because they had the largest training sets *token*-wise, whereas Turkish had the smallest training set and lowest F_1 scores.

5.2 Monolingual selectivity

While the diagnostic probes drastically outperformed their controlled counterparts, we do see a trend of selectivity improving with the number of layers. This reinforces the findings of Hewitt and Liang (2019), who posit that classifiers trained on top of lower layers are better equipped to memorize input-output mappings, due to their proximity to the initial vocabulary representations of the embedding layer. Nevertheless, the high selectivity scores across the probes show that a multilabel probing classifier offers a promising diagnostic of morphosyntactic representations.

From a cross-linguistic standpoint, it is interesting that the probes for Afrikaans—the one morphologically *impoverished* language in the bunch—exhibited the worst selectivity. This suggests that, perhaps, it is easier for probes to memorize mappings for analytic languages (i.e., languages that lack rich inflectional systems). However, as the the Afrikaans probes were trained on the second largest number of tokens, they may have had more opportunity to memorize the control task. (Similarly, the Spanish probes, which had the largest training set, displayed the second best performance on the control task.)

5.3 Case study: Hebrew covert determiners

The micro-averaged scores in Figure 2 show that mBERT has indeed learned *some* linguistic system or portion thereof. However, these scores do not give much insight into which aspects of morphosyntax mBERT has come to represent, the interplay between these properties, nor how much mBERT varies in capturing each feature value. Crucially, a key strength of multilabel probing is that it makes it easy to mine fine-grained morphosyntactic observations that implicate multiple features. In this section, we present such an analysis with Hebrew determiners, inspired by Klein and Tsarfaty (2020). We focus on the predictions from mBERT-6, since it displayed the highest F_1 and selectivity scores out of the Hebrew probes.

Ambiguous orthographies as well as multiword tokens (MWTs) are ubiquitous in Hebrew. As stated previously, we represented MWTs by flattening their structure and labeling each MWT with the feature labels of its components. A common structure of MWTs in Hebrew is ADP-(DET)-NOUN, where the determiner is the definite article -ה *ha* 'the'. Depending on the preposition, the definite article is represented orthographically (e.g., -מה *miha* 'from the') or as a vowel change on the preposition that is not represented orthographically (e.g., -ל can be either *le* 'to a' or *la* 'to the'). When the article is absent from the orthography, we refer to it as being *covert*.[4]

The definite article is one type of determiner in the HTB corpus, but is uniquely identified by the label `PronType=Art`. We thus extracted all of the ADP-(DET)-NOUN cases from the Hebrew test set (234 in total) and examined how well mBERT-6 captured this property. We found that it was less able to recognize `PronType=Art` when the article was not overt (Table 1).

Yet, we also found that agreement patterns facilitated recognition of the covert definite article. In particular, Hebrew adjectival modifiers agree with the nouns they modify in gender, number, and definiteness (e.g., in the noun phrase הבית הקטן *habayit hakatan* 'the small house', בית *bayit* is 'house.sing.masc', קטן *katan* is 'small.sing.masc', and -ה *ha* is the definite article). Based on UD's `amod` annotations, the MWTs that appeared in these constructions constituted 44.3% of TPs, 19.4% of FPs, and 26.2% of FNs when identifying the covert definite article. Moreover, the majority of the FNs involved additional erroneous predictions, where either `PronType=Art` was not captured on the modifier, the parts of speech were misidentified, or the modifier and the noun were mis-predicted to disagree along an additional feature (i.e., gender or number). These concomitant errors were largely missing from the TPs.

It seems that mBERT-6 has learned that Hebrew nouns and their modifiers agree along multiple features, and that it is able to use the presence of an overt definite article on a modifier to help infer the presence of a covert article in a MWT. When not all of the grammatical features that participate in agreement are captured, this can attenuate recognition of the covert article (and vice versa).

6 Multilingual experiments

We have used monolingual probes to assess the linguistic representations from mBERT on a language-by-language basis. However, can we replace the individual monolingual probes with a single multilingual probe and derive comparable insights? To address this question, we trained multilingual probes on a shuffled combination of the training sets for Afrikaans, Croatian, Finnish, Hebrew, Spanish, and Turkish. The multilingual probes extracted an aggregated subset of the features captured by the monolingual probes. We then assessed the multilingual probes' performance on each language independently. Overall, the multilingual probes exhibited slight dips in performance, but better selectivity, compared to their monolingual counterparts (Figure 3). These trends occurred despite all of the multilingual models converging before they reached epoch 50.

`PronType=Art`	P	R	F_1
Overt determiner	0.93	0.56	0.70
Covert determiner	0.69	0.40	0.50

Table 1: Recognition of the feature `PronType=Art` in ADP-DET-NOUN multiword tokens, given the Hebrew mBERT-6 probe.

Probe	Af	Hr	Fi	Es	Tr
Mono.	0.89	0.88	0.89	0.96	0.79
Multi.	0.71	0.76	0.80	0.14	0.65

Table 2: F_1 results for nominative case (`Case=Nom`) in Afrikaans (Af), Croatian (Hr), Finnish (Fi), Spanish (Es), and Turkish (Tr), given the monolingual and multilingual mBERT-6 probes.

6.1 Multilingual task complexity

Even though the multilingual experiments merely combine the monolingual training data, the multilingual task is inherently more complex than the monolingual task. Namely, the probes must balance the needs of multiple languages and extract features from a broader diversity of data.

Let us consider nominative case. When focusing on predictions from mBERT-6, we see that the `Case=Nom` scores for each language dipped with the multilingual probe (Table 2). Importantly, the distribution of nominative morphology differs cross-linguistically; according to the UD corpora, for instance, nominative inflections appear on nouns, verbs, and adjectives in Turkish, but only on pronouns in Spanish. It is possible that such variation might result in "conflicting" train-

[4] Since the article is (optionally) audible, this usage of *covert* differs slightly from its usage in linguistic theory.

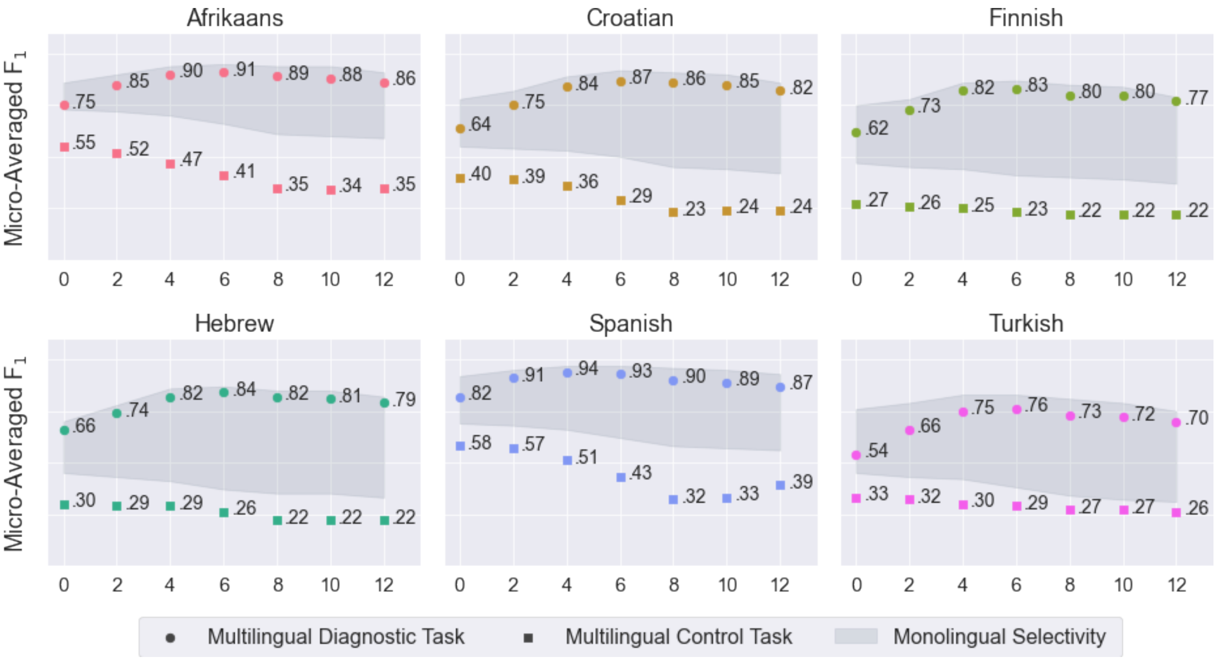


Figure 3: Micro-averaged F_1 results from the multilingual probes on the diagnostic and control tasks for each language. The x-axes indicate the mBERT layer. The depicted monolingual results (for comparison) assume the same feature label subsets as the multilingual models; incidentally, the monolingual diagnostic task scores are equivalent to the scores reported in Figure 2, while the control task scores differ by ±2 points.

ing signals to the probe, causing the performance of the multilingual probes to dip. Furthermore, it suggests that, although mBERT renders nominative case easily extractable for each language independently, mBERT has not recognized their nominative morphology to correspond to the same nominative notion. We return to this point in §7.

6.2 Hints of memorization

Indeed, another potential explanation for the contrast in monolingual and multilingual performance is that the simpler task affords the monolingual probes more opportunity to memorize the feature labels. This explanation, which is explored further in Appendix D, is supported by how the multilingual probes generally exhibit greater selectivity and accounts for why their performance deficit is, for the most part, spread evenly across the feature labels (see Appendix E for the full feature-level results).

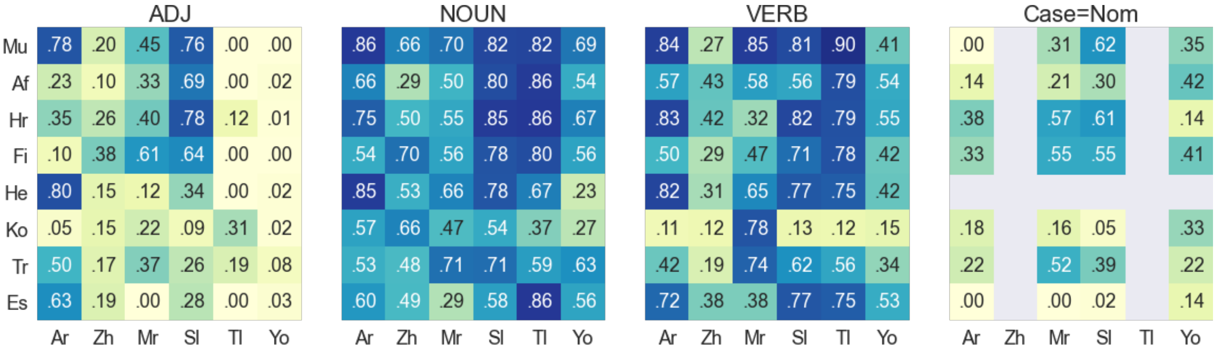
7 Crosslingual experiments

Our probing paradigm can also be used to study which morphosyntactic features are encoded similarly cross-linguistically: If a monolingual probe can successfully extract a feature label given a *held-out* language, this suggests that the LM has come to recognize that property as being shared by the two languages.

In this section, we evaluate the monolingual and multilingual probes on UD test sets for Arabic, Chinese, Marathi, Slovenian, Tagalog, and Yorùbá. These experiments are akin to prior work on zero-shot crosslingual transfer (Pires et al., 2019; Wu and Dredze, 2019; Conneau et al., 2020b; K et al., 2020), though we differ in that we never fine-tune mBERT. Focusing once more on mBERT-6, this section examines a small subset of labels, presented in Figure 4. However, see Appendix F for the global F_1 scores across the held-out languages and full feature-level results from mBERT-6.

7.1 Towards cross-linguistic categories

Overall, the probes performed relatively well on extracting nouns and verbs across the held-out languages. This suggests that mBERT encodes *noun*-hood and *verb*-hood in a cross-linguistic fashion—that it has some conception of nouns and verbs that transcends individual languages. *Adjective*-hood, in contrast, seems to be represented less cohesively. The probes struggled to identify adjectives in Chinese, and even more so in Tagalog and Yorùbá. This is not to say that mBERT does not capture adjectives in these languages, but, rather, that it has not connected them to their counterparts in other



ADJ	Ar	Zh	Mr	Sl	Tl	Yo
Mu	.78	.20	.45	.76	.00	.00
Af	.23	.10	.33	.69	.00	.02
Hr	.35	.26	.40	.78	.12	.01
Fi	.10	.38	.61	.64	.00	.00
He	.80	.15	.12	.34	.00	.02
Ko	.05	.15	.22	.09	.31	.02
Tr	.50	.17	.37	.26	.19	.08
Es	.63	.19	.00	.28	.00	.03

NOUN	Ar	Zh	Mr	Sl	Tl	Yo
Mu	.86	.66	.70	.82	.82	.69
Af	.66	.29	.50	.80	.86	.54
Hr	.75	.50	.55	.85	.86	.67
Fi	.54	.70	.56	.78	.80	.56
He	.85	.53	.66	.78	.67	.23
Ko	.57	.66	.47	.54	.37	.27
Tr	.53	.48	.71	.71	.59	.63
Es	.60	.49	.29	.58	.86	.56

VERB	Ar	Zh	Mr	Sl	Tl	Yo
Mu	.84	.27	.85	.81	.90	.41
Af	.57	.43	.58	.56	.79	.54
Hr	.83	.42	.32	.82	.79	.55
Fi	.50	.29	.47	.71	.78	.42
He	.82	.31	.65	.77	.75	.42
Ko	.11	.12	.78	.13	.12	.15
Tr	.42	.19	.74	.62	.56	.34
Es	.72	.38	.38	.77	.75	.53

Case=Nom	Ar	Zh	Mr	Sl	Tl	Yo
Mu	.00		.31	.62		.35
Af	.14		.21	.30		.42
Hr	.38		.57	.61		.14
Fi	.33		.55	.55		.41
He						
Ko	.18		.16	.05		.33
Tr	.22		.52	.39		.22
Es	.00		.00	.02		.14

Figure 4: A handful of feature-level F$_1$ results from evaluating the monolingual and multilingual mBERT-6 probes on "held-out" languages. The x-axes indicate the held-out language (Ar=Arabic, Zh=Chinese, Mr=Marathi, Sl=Slovenian, Tl=Tagalog, and Yo=Yorùbá), while the y-axes indicate the probe (Mu=Multilingual, Af=Afrikaans, Hr=Croatian, Fi=Finnish, He=Hebrew, Ko=Korean, Es=Spanish, and Tr=Turkish). Grayed-out regions indicate where the feature label is not applicable to the language or annotated in the language's corpus.

languages. This may be especially true for low-resource languages like Tagalog and Yorùbá.[5] Even though mBERT's training involved over-sampling smaller corpora, it might be the case that the model required exposure to Tagalog and Yorùbá adjectives in a wider array of contexts in order to relate them to their counterparts cross-linguistically (see Conneau et al. 2020a for interesting discussion).

Cross-linguistic variation in a feature's distribution in natural languages might also lead a LM not to recognize when a property is shared by multiple languages. In §6, we cited such variation as the reason the multilingual probes struggled with nominative case. We see this suspicion further borne out in Figure 4, where predictions of `Case=Nom` in the held-out languages ranged from 0 to 0.62 F$_1$. As evidenced by this lack of transfer, it seems that cross-linguistic variation in the distribution of nominative morphology led to a decentralized encoding of nominative case in mBERT; consequently, this made it more challenging for the probes to capture nominative case in the held-out languages (and for the multilingual probes to identify nominative case in general).

Yet, there are also cases where the multilingual probes performed better than the monolingual probes with the held-out languages. Most strikingly, the mBERT-6 multilingual probe obtained 0.90 F$_1$ on Tagalog verbs, whereas none of the monolingual probes got over 0.79 F$_1$. This suggests that, with cross-linguistic properties that are encoded more cohesively, such as *verb*-hood, exposure to multiple languages can lead a probe to forge more replete connections with mBERT's representational space.

7.2 Family ties

In the absence of cross-linguistic representations, we generally find that a monolingual probe extends equivalently or better to a held-out language than the multilingual model. In particular, the monolingual probes often did well with *related* languages (cf. Pires et al., 2019; Wu and Dredze, 2019; Conneau et al., 2020b). Compared to the other monolingual probes, for instance, the Hebrew probes fared best with Arabic, another Semitic language, topping out at a micro-averaged F$_1$ score of 0.56 (see Appendix F). This was also the case at the feature level with nouns, verbs, and adjectives, as shown in Figure 4. Notably, Hebrew and Arabic use different scripts. If mBERT has come to represent them similarly, this likely falls out of the structural similarities between the two languages.

Likewise, the Croatian mBERT-6 probe achieved a micro-averaged F$_1$ score of 0.70 on Slovenian. (For comparison, the Turkish mBERT-6 probe scored 0.76 F$_1$ *on Turkish.*) The Croatian probe also performed the best on Slovenian nouns, verbs, and adjectives, as well as with words inflected for first person, plurality, or indicative mood. This success seems due to both structural and surface similarities (e.g., cognates) between Croatian and Slovenian. For example, Croatian achieved 0.95 F$_1$ on conditional mood (`Mood=Cnd`; see Appendix

[5] Recall that mBERT was trained on the 100 languages with the largest Wikipedias. Based on Wikimedia's *List of Wikipedias*, it seems that the Wikipedia dumps for Tagalog and Yorùbá were among the smallest corpora that mBERT was trained on, ranking 92 and 106 at present, respectively. Note also that, in the "language resource race", Joshi et al. (2020) give Tagalog and Yorùbá scores of 3/5 and 2/5.

F) and 0.86 F_1 on indicative mood (`Mood=Ind`) in Slovenian because the two languages share several auxiliaries that mark mood (e.g., *bi* for conditional, *je* for indicative).

7.3 Revisiting memorization

Note that, with the exception of shared morphemes, the successful instances of crosslingual transfer cannot be reduced to memorization. If the probes merely memorized their monolingual training data, one would expect chance performance and less variability when evaluating them on the held-out languages. These evaluations further verify that the multilabel probes extracted meaningful representations from mBERT. When applied to held-out languages, they also provide a supplementary method for gauging the complexity of a probe and its ability to memorize a linguistic task.

8 Discussion and conclusion

Emerging studies on interpretability have highlighted a wealth of linguistic information that can be extracted from neural language models. Contributing to this effort, we propose using a multilabel probing task to analyze the morphosyntactic representations of multilingual word embeddings. We demonstrate this probing paradigm with mBERT (Devlin et al., 2018).

In a set of monolingual experiments (§5), we trained individual probes for Afrikaans, Croatian, Finnish, Hebrew, Korean, Spanish, and Turkish. We found that mBERT-6 holds the most morphosyntactic information (cf. Liu et al., 2019; Tenney et al., 2019), with the probes obtaining micro-averaged F_1 scores between 0.83 and 0.97. In a small case study of Hebrew determiners (§5.3), we illustrated an analysis that implicates *multiple* features (i.e., lexical category, pronominal type, number, and gender). Crucially, traditional single-label efforts would require training multiple models to arrive at such an analysis and, in general, run the risk of overlooking relevant features. (We also suspect that training multiple one-off probes is less computationally efficient than training a single multilabel probe, though we leave this comparison for future work.)

Next, in a set of multilingual experiments (§6), we saw that the multilingual probes marginally underperformed their monolingual counterparts, while largely upholding the same trends and exhibiting better selectivity. We attributed this contrast in performance to the monolingual probes relying more on memorization, given a simpler task (§6.2). These findings indicate that the multilingual probes may be more "expressive" diagnostics of linguistic representations (cf. Hewitt and Liang, 2019). However, since our goal is to probe embeddings rather than to perform state-of-the-art morphosyntactic tagging, the monolingual and multilingual probes offer the same *insights* to the extent that they exhibit comparable trends and lend themselves to the same generalizations.

In a set of crosslingual experiments, we further evaluated the monolingual and multilingual probes on data from six "held-out" languages: Arabic, Chinese, Marathi, Slovenian, Tagalog, and Yorùbá (§7). We showed that applying the probes accordingly can help illuminate which linguistic properties a LM recognizes as being shared by multiple languages and what factors might lead a LM not to encode cohesive representations of a particular cross-linguistic feature. Namely, we conjectured that cross-linguistic variation in the distribution of nominative morphology led mBERT to form decentralized representations of nominative case; in turn, this made it more challenging for the probes to extract nominative case in the held-out languages.

In sum, multilabel probe predictions can be used to perform holistic analyses of a language model's ability to encode systems of morphology, as well as more fine-grained analyses of individual features, agreement phenomena, and how shared properties are represented cross-linguistically. We release the predictions from our probes to support more detailed analyses of mBERT's facility for morphosyntax; these predictions can also be used to focus future contributions by identifying which mBERT layers to target for more complex probing of specific features. In addition, we encourage future efforts to probe different multilingual language models using the multilabel paradigm and to examine how these models might vary in their morphosyntactic representations (cf. Mikhailov et al., 2021). Finally, future research should explore how global and feature-level morphosyntactic probe performance corresponds to the performance of downstream systems, especially amongst morphologically rich languages.

Acknowledgments

We thank Fei Xia, the BlackboxNLP 2020 attendees, and the EMNLP 2021 reviewers for their helpful feedback and discussions. We also thank the UW Research Computing Club for supporting our research through their Cloud Credit Program.

Ethical considerations

While our proposed probing paradigm is intended for analyzing large pre-trained language models, which are computationally (and monetarily) expensive to produce (cf. Strubell et al., 2019; Bender et al., 2021), our probes are lightweight and quick to train. To help minimize our use of computational resources, we deployed a "cache and batch" approach to pre-processing our data, which we describe in Appendix C. Furthermore, in addition to releasing our code, we share our multilabel probe predictions to facilitate future morphosyntactic analyses of mBERT (i.e., without the need for training analogous probes).

In our experiments, we prioritized working with data from a typologically diverse set of languages, many of which are understudied in the field of natural language processing (cf. Joshi et al., 2020). In particular, we drew on data from Universal Dependencies (Nivre et al., 2016; Dobrovoljc and Nivre, 2016), working with morphologically-annoated corpora for 13 different languages: Afrikaans (AfriBooms; cf. Dirix et al., 2017), Arabic (PADT; cf. Smrž et al., 2002, 2008; Hajič et al., 2009), Chinese (PUD; cf. Zeman et al., 2017), Croatian (SET; cf. Agić and Ljubešić, 2015), Finnish (TDT; cf. Haverinen et al., 2014; Pyysalo et al., 2015), Hebrew (HTB; cf. Tsarfaty, 2013; McDonald et al., 2013; Sadde et al., 2018), Korean (PUD; cf. Zeman et al., 2017), Marathi (UFAL; cf. Ravishankar, 2017), Slovenian (SST; cf. Dobrovoljc and Nivre, 2016), Spanish (AnCora; cf. Alonso and Zeman, 2016), Tagalog (TRG), Turkish (IMST; cf. Sulubacak et al., 2016; Tyers et al., 2017; Türk et al., 2019), and Yorùbá (YTB; cf. Ishola and Zeman, 2020). Appendix A briefly summarizes the subsets of these datasets that we used in our experiments.

Though Universal Dependencies is an incredible resource—rich with morphosyntactic and dependency annotations—it is important to remember that many of the these datasets source texts from somewhat narrow domains (e.g., Wikipedia, news corpora, Bible passages) and, thus, may be limited in the linguistic phenomena they capture. Moreover, these datasets are accompanied by varying degrees of documentation. Please see our repository for further details about these datasets (in the form of Bender and Friedman-inspired data statements) and for a more thorough discussion of the ethical considerations relevant to our paper.[6]

References

Željko Agić and Nikola Ljubešić. 2015. Universal Dependencies for Croatian (that work for Serbian, too). In *Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing*, pages 1–8.

Guillaume Alain and Yoshua Bengio. 2018. Understanding intermediate layers using linear classifier probes. *arXiv:1610.01644*.

Hector Martinez Alonso and Daniel Zeman. 2016. Universal Dependencies for the AnCora treebanks. In *Procesamiento del Lenguaje Natural, Sociedad Espanola para el Procesamiento del Lenguaje Natural*, pages 91–98.

Geoff Bacon and Terry Regier. 2019. Does BERT agree? Evaluating knowledge of structure dependence through agreement relations. *arXiv:1908.09892*.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017a. 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.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017b. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the 8th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, page 610–623.

Terra Blevins and Luke Zettlemoyer. 2019. Better character language modeling through morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1606–1613.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the*

[6] https://github.com/tsnaomi/morph-bert

58th Annual Meeting of the Association for Computational Linguistics, pages 5564–5577.

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

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#*vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136.

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.

Jacob Devlin, Min-Wei Chang, Kenton Lee, and Toutanova Kristina. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*.

Peter Dirix, Liesbeth Augustinus, Daniel van Niekerk, and Frank Van Eynde. 2017. Universal Dependencies for Afrikaans. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 38–47.

Kaja Dobrovoljc and Joakim Nivre. 2016. The Universal Dependencies treebank of spoken Slovenian. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 1566–1573.

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4423–4437.

Daniel Edmiston. 2020. A systematic analysis of morphological content in BERT models for multiple languages. *arXiv:2004.03032*.

Richard Futrell and Roger P. Levy. 2019. Do RNNs learn human-like abstract word order preferences? In *Proceedings of the Society for Computation in Linguistics*, pages 50–59.

Daniela Gerz, Ivan Vulić, Edoardo Ponti, Jason Naradowsky, Roi Reichard, and Anna Korhonen. 2018. Language modeling for morphologically rich languages: Character-aware modeling for word-level prediction. *Transactions of the Association for Computational Linguistics*, 6:451–465.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205.

Jan Hajič, Otakar Smrž, Petr Zemánek, Petr Pajas, Jan Šnaidauf, Emanuel Beška, Jakub Kracmar, and Kamila Hassanová. 2009. Prague Arabic dependency treebank 1.0. Technical report.

Katri Haverinen, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Anna Missilä, Stina Ojala, Tapio Salakoski, and Filip Ginter. 2014. Building the essential resources for Finnish: The Turku Dependency Treebank. *Language Resources and Evaluation*, 48(3):493–531.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing)*.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. 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)*.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks rocess hierarchical structure. *Journal of Artificial Intelligence Research*, 61(1):907–926.

Ìájídé Ishola and Daniel Zeman. 2020. Yorùbá Dependency Treebank (YTB). In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5178–5186.

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

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

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

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *arXiv:1412.6980*.

Stav Klein and Reut Tsarfaty. 2020. Getting the ##life out of living: How adequate are word-pieces for modelling complex morphology? In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 204–209. Association for Computational Linguistics.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. 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 1073–1094.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.

Arya D McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J Mielke, Jeffrey Heinz, et al. 2019. The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244.

Ryan T McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith B Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 2013. Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97.

Vladislav Mikhailov, Oleg Serikov, and Ekaterina Artemova. 2021. Morph Call: Porbing morphosyntactic content of multilingual transformers. In *Proceedings of the 3rd Workshop on Computational Typology and Multilingual NLP*, pages 97–121.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 1659–1666. European Language Resources Association (ELRA).

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043.

Hyunji Hayley Park, Katherine J Zhang, Coleman Haley, Kenneth Steimel, Han Liu, and Lane Schwartz. 2020. Morphology matters: A multilingual language modeling analysis. *arXiv:2012.06262*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035.

Tiago Pimentel, Naomi Saphra, Adina Williams, and Ryan Cotterell. 2020a. Pareto probing: Trading off accuracy for complexity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 3138–3153. Association for Computational Linguistics.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020b. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Sampo Pyysalo, Jenna Kanerva, Anna Missilä, Veronika Laippala, and Filip Ginter. 2015. Universal Dependencies for Finnish. In *Proceedings of the 20th Nordic Conference of Computational Linguistics*, pages 163–172.

Vinit Ravishankar. 2017. A Universal Dependencies Treebank for Marathi. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 190–200.

Shoval Sadde, Amit Seker, and Reut Tsarfaty. 2018. The Hebrew Universal Dependency Treebank: Past, present and future. In *Proceedings of the 2nd Workshop on Universal Dependencies*, pages 133–143.

Gözde Gül Şahin, Clara Vania, Ilia Kuznetsov, and Iryna Gurevych. 2020. LINSPECTOR: Multilingual probing tasks for word representations. *Computational Linguistics*, 46(2):335–385.

Otakar Smrž, Viktor Bielický, Iveta Kouřilová, Jakub Kráčmar, Jan Hajič, and Petr Zemánek. 2008. Prague Arabic Dependency Treebank: A word on the million words. In *Proceedings of the Workshop on Arabic and Local Languages*, pages 16–23.

Otakar Smrž, Jan Šnaidauf, and Petr Zemánek. 2002. Prague Dependency Treebank for Arabic: Multi-level annotation of Arabic corpus. In *Proceedings of the International Symposium on the Processing of Arabic*, pages 147–155.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650.

Umut Sulubacak, Memduh Gökirmak, Francis M Tyers, Çağri Çöltekin, Joakim Nivre, and Gülşen Eryiğit. 2016. Universal Dependencies for Turkish. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3444–3454.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.

Reut Tsarfaty. 2013. A unified morpho-syntactic scheme of Stanford Dependencies. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Utku Türk, Furkan Atmaca, Şaziye Betül Özateş, Balkız Öztürk Başaran, Tunga Güngör, and Arzucan Özgür. 2019. Improving the annotations in the Turkish Universal Dependency Treebank. In *Proceedings of the 3rd Workshop on Universal Dependencies*.

Francis M Tyers, Jonathan Washington, Çağrı Çöltekin, and Aibek Makazhanov. 2017. An assessment of Universal Dependency annotation guidelines for Turkic languages. In *Proceedings of the 5th International Conference on Turkic Language Processing*, pages 276–297.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, Alexander M. RushThomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

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*, pages 833–844.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19.

Kelly W. Zhang and Samuel R. Bowman. 2018. Language modeling teaches you more syntax than translation does: Lessons learned through auxiliary task analysis. *arXiv:1809.10040*.

Table A1: Composition of the training and evaluation data for the monolingual and multilingual probes.

Language	Genus	$\|F\|$	Train		Dev		Test	
			Sentences	Tokens	Sentences	Tokens	Sentences	Tokens
Afrikaans	Germanic	53	800	21,160	194	5,317	425	10,065
Croatian	Slavic	66	800	17,811	960	22,292	1,136	24,260
Finnish	Finnic	89	800	10,786	1,363	18,311	1,553	21,069
Hebrew	Semitic	53	800	16,061	484	8,358	491	8,829
Korean	Korean	35	800	13,177	100	1,679	100	1,728
Spanish	Romance	63	800	24,345	1,654	52,161	1,719	52,429
Turkish	Turkic	64	800	8,244	983	9,768	981	9,794
Multilingual	n/a	72	4,800	98,297	5,638	116,207	n/a	n/a

Table A2: Composition of the "held-out" language data (GCP = Greater Central Philippine).

Language	Genus	Test	
		Sentences	Tokens
Arabic	Semitic	675	24,195
Chinese	Chinese	1,000	21,415
Korean	Korean	1,000	16,584
Marathi	Indic	47	376
Slovenian	Slavic	995	9,880
Tagalog	GCP	55	292
Yorùbá	Defoid	318	8,198

Appendix A Universal Dependencies

We performed multilabel probing using morphologically annotated corpora from Universal Dependencies (UD). Table A1 summarizes the datasets for the monolingual and multilingual experiments and Table A2 for the crosslingual experiments.

Appendix B Feature labels

Tables B1 and B2 list the 166 feature labels we extracted in total across our experiments. The monolingual probes were trained to extract every morphosyntactically relevant label that was available for a given language in its UD corpus. The multilingual probes focused on a subset of these labels.

Appendix C Implementation details

Word-level predictions To perform word-level predictions of morphosyntactic properties, we first passed the raw corpus sentences through mBERT, then aggregated the contextualized word embeddings on a word-by-word basis. In small exploratory experiments, we found that summing the subword embeddings performed the best; we thus used this aggregation strategy throughout our experiments. Notably, summing the subword representations achieved comparable F_1 scores but higher selectivity than taking their average. The summation and averaging strategies also performed better than representing each word by the embedding for its word-initial or word-final word piece.

Cache and batch Prior to training, we cached the aggregated word representations; these stored embeddings then served as inputs to the probes. This was done in lieu of passing a batch of input sentences through mBERT and doing the aggregation on the fly at each training step. Since the probes themselves are simple linear layers and therefore non-contextual, we were able to batch the embeddings at the token level: We dispensed with the sequence length dimension and skipped padding. In all of the experiments, we opted for a batch size of 512 tokens (i.e., the batches had a dimensionality of 512×768). This "cache and batch" approach allowed each monolingual probe to train in $\sim$1 minute and each multilingual probe in $\sim$4 minutes on a Tesla K80 GPU.

Control task In the control task, each word type $v_i \in V$ was assigned a multi-hot output vector $\mathbf{c^i}$, where c^i_j was sampled according to the true distribution of the feature label f_j in the training data.[7] To help ensure the presence of controlled counterparts for low-frequency feature labels, each feature label had a minimum probability threshold of 0.001. For each language, the probes for the various mBERT layers were trained to predict the same set of random output vectors.

[7] V is based on the word types across the training, validation, and test sets, since UD corpora use an open vocabulary.

Table B1: The monolingual probes extracted different sets of features, while the multilingual probes extracted a semi-aggregated subset of these features (in bold under "Feature Labels").

Feature Labels	Afrikaans	Croatian	Finnish	Hebrew	Korean	Spanish	Turkish
ADJ	✓	✓	✓	✓	✓	✓	✓
ADP	✓	✓	✓	✓		✓	✓
ADV	✓	✓	✓	✓	✓	✓	✓
AUX	✓	✓	✓	✓	✓	✓	✓
CCONJ	✓	✓	✓	✓	✓	✓	✓
DET	✓	✓		✓	✓	✓	✓
NOUN	✓	✓	✓	✓	✓	✓	✓
NUM	✓	✓	✓	✓	✓	✓	✓
PART	✓	✓			✓	✓	
PRON	✓	✓	✓	✓	✓	✓	✓
PROPN	✓	✓	✓	✓	✓	✓	✓
SCONJ	✓	✓	✓	✓		✓	
VERB	✓	✓	✓	✓	✓	✓	✓
AdjType=Attr	✓						
AdjType=Pred	✓						
AdpType=Post			✓				
AdpType=Prep	✓		✓			✓	
AdpType=Preppron						✓	
AdvType=Tim						✓	
Animacy=Anim		✓					
Animacy=Inan		✓					
Aspect=Hab							✓
Aspect=Perf							✓
Aspect=Prog							✓
Aspect=Prosp							✓
Aspect=Rapid							✓
Case=Abe			✓				
Case=Abl			✓				✓
Case=Acc	✓	✓	✓	✓	✓	✓	✓
Case=Ade			✓				
Case=Advb					✓		
Case=All			✓				
Case=Com			✓			✓	
Case=Comp					✓		
Case=Dat		✓				✓	✓
Case=Ela			✓				
Case=Equ							✓
Case=Ess			✓				
Case=Gen		✓	✓	✓	✓		✓
Case=Ill			✓				
Case=Ine			✓				
Case=Ins		✓	✓				✓
Case=Loc		✓					✓
Case=Nom	✓	✓	✓		✓	✓	✓
Case=Par			✓				
Case=Tem				✓			
Case=Tra			✓				

Continuation of Table B1:

Feature Labels	Afrikaans	Croatian	Finnish	Hebrew	Korean	Spanish	Turkish
Case=Voc		✓					
Clitic=Han			✓				
Clitic=Ka			✓				
Clitic=Kaan			✓				
Clitic=Kin			✓				
Clitic=Ko			✓				
Clitic=Pa			✓				
Clitic=S			✓				
Connegative=Yes			✓				
Definite=Cons				✓			
Definite=Def	✓	✓		✓		✓	
Definite=Ind	✓	✓				✓	
Degree=Abs						✓	
Degree=Cmp	✓	✓	✓			✓	
Degree=Dim	✓						
Degree=Pos	✓	✓	✓				
Degree=Sup	✓	✓	✓			✓	
Derivation=Inen			✓				
Derivation=Ja			✓				
Derivation=Lainen			✓				
Derivation=Llinen			✓				
Derivation=Minen			✓				
Derivation=Sti			✓				
Derivation=Tar			✓				
Derivation=Ton			✓				
Derivation=Ttain			✓				
Derivation=U			✓				
Derivation=Vs			✓				
Echo=Rdp							✓
Evident=Nfh							✓
Form=Adn					✓		
Form=Aux					✓		
Form=Compl					✓		
Gender=Fem		✓		✓		✓	
Gender=Masc		✓		✓		✓	
Gender=Neut		✓					
Gender[psor]=Fem		✓					
Gender[psor]=Masc		✓					
Gender[psor]=Neut		✓					
HebBinyan=HIFIL				✓			
HebBinyan=HITPAEL				✓			
HebBinyan=HUFAL				✓			
HebBinyan=NIFAL				✓			
HebBinyan=PAAL				✓			
HebBinyan=PIEL				✓			
HebBinyan=PUAL				✓			
HebExistential=True				✓			

Continuation of Table B1:

Feature Labels	Afrikaans	Croatian	Finnish	Hebrew	Korean	Spanish	Turkish
InfForm=1			✓				
InfForm=2			✓				
InfForm=3			✓				
Mood=Cnd		✓	✓			✓	✓
Mood=Des							✓
Mood=Gen							✓
Mood=Imp		✓	✓	✓	✓	✓	✓
Mood=Ind		✓	✓		✓	✓	✓
Mood=Nec							✓
Mood=Opt							✓
Mood=Pot			✓				✓
Mood=Sub						✓	
NumType=Card		✓	✓		✓	✓	✓
NumType=Dist							✓
NumType=Frac						✓	
NumType=Mult		✓					
NumType=Ord		✓	✓			✓	✓
Number=Dual							
Number=Plur	✓	✓	✓	✓	✓	✓	✓
Number=Sing	✓	✓	✓	✓		✓	✓
Number[psor]=Plur		✓	✓			✓	✓
Number[psor]=Sing		✓	✓			✓	✓
PartForm=Agt			✓				
PartForm=Neg							
PartForm=Past			✓				
PartForm=Pres			✓				
PartType=Gen	✓						
PartType=Inf	✓						
PartType=Neg	✓						
Person=0			✓				
Person=1	✓	✓	✓	✓	✓	✓	✓
Person=2	✓	✓	✓	✓	✓	✓	✓
Person=3	✓	✓	✓	✓	✓	✓	✓
Person[psor]=1			✓				✓
Person[psor]=2			✓				✓
Person[psor]=3			✓				✓
Polarity=Neg		✓	✓	✓	✓	✓	✓
Polarity=Pos				✓			✓
Polite=Form					✓	✓	✓
Polite=Infm							✓
Poss=Yes	✓	✓				✓	
Prefix=Yes				✓			
PrepCase=Npr						✓	
PrepCase=Pre						✓	
PronType=Art	✓			✓		✓	
PronType=Dem	✓	✓	✓	✓		✓	✓
PronType=Emp				✓			

Continuation of Table B1:

Feature Labels	Afrikaans	Croatian	Finnish	Hebrew	Korean	Spanish	Turkish
PronType=Ind	✓	✓	✓	✓		✓	✓
PronType=Int	✓	✓	✓	✓	✓	✓	
PronType=Neg		✓				✓	
PronType=Prs	✓	✓	✓	✓		✓	✓
PronType=Rcp			✓				
PronType=Rel	✓	✓	✓			✓	
PronType=Tot		✓				✓	
Reflex=Yes	✓	✓	✓	✓		✓	✓
Subcat=Intr	✓						
Subcat=Prep	✓						
Subcat=Tran	✓						
Tense=Fut				✓	✓	✓	✓
Tense=Imp		✓				✓	
Tense=Past	✓	✓	✓	✓	✓	✓	✓
Tense=Pqp							✓
Tense=Pres	✓	✓	✓			✓	✓
VerbForm=Conv		✓					✓
VerbForm=Fin	✓	✓	✓		✓	✓	
VerbForm=Ger					✓	✓	
VerbForm=Inf	✓	✓	✓	✓		✓	
VerbForm=Part	✓	✓	✓	✓		✓	✓
VerbForm=Vnoun							✓
VerbType=Aux	✓						
VerbType=Cop	✓			✓			
VerbType=Mod	✓			✓			
VerbType=Pas	✓						
Voice=Act		✓	✓	✓			
Voice=Cau					✓		✓
Voice=Mid				✓			
Voice=Pass		✓	✓	✓	✓		✓

Table B2: The monolingual and multilingual probes were evaluated on seven “held-out” languages.

Feature Labels	Arabic	Chinese	Marathi	Slovenian	Tagalog	Yorùbá
ADJ	✓	✓	✓	✓	✓	✓
ADP	✓	✓	✓	✓	✓	✓
ADV	✓	✓	✓	✓	✓	✓
AUX	✓	✓	✓	✓	✓	✓
CCONJ	✓	✓	✓	✓		✓
DET	✓	✓	✓	✓	✓	✓
NOUN	✓	✓	✓	✓	✓	✓
NUM	✓	✓	✓	✓		✓
PART	✓	✓	✓	✓	✓	✓
PRON	✓	✓	✓	✓	✓	✓
PROPN	✓	✓	✓	✓	✓	✓
SCONJ	✓	✓	✓	✓	✓	✓
VERB	✓	✓	✓	✓	✓	✓
AdjType=Attr						
AdjType=Pred						
AdpType=Post						
AdpType=Prep						
AdpType=Preppron						
AdvType=Tim						
Animacy=Anim						
Animacy=Inan						
Aspect=Hab						
Aspect=Perf						
Aspect=Prog						
Aspect=Prosp						
Aspect=Rapid						
Case=Abe						
Case=Abl						
Case=Acc	✓		✓	✓		✓
Case=Ade						
Case=Advb						
Case=All						
Case=Com						
Case=Comp						
Case=Dat			✓	✓	✓	
Case=Ela						
Case=Equ						
Case=Ess						
Case=Gen	✓	✓		✓		✓
Case=Ill						
Case=Ine						
Case=Ins			✓	✓		
Case=Loc			✓	✓	✓	
Case=Nom	✓		✓	✓		✓
Case=Par						
Case=Tem						
Case=Tra						

Continuation of Table B2:

Feature Labels	Arabic	Chinese	Marathi	Slovenian	Tagalog	Yorùbá
Case=Voc			✓			
Clitic=Han						
Clitic=Ka						
Clitic=Kaan						
Clitic=Kin						
Clitic=Ko						
Clitic=Pa						
Clitic=S						
Connegative=Yes						
Definite=Cons						
Definite=Def						
Definite=Ind						
Degree=Abs						
Degree=Cmp						
Degree=Dim						
Degree=Pos						
Degree=Sup						
Derivation=Inen						
Derivation=Ja						
Derivation=Lainen						
Derivation=Llinen						
Derivation=Minen						
Derivation=Sti						
Derivation=Tar						
Derivation=Ton						
Derivation=Ttain						
Derivation=U						
Derivation=Vs						
Echo=Rdp						
Evident=Nfh						
Form=Adn						
Form=Aux						
Form=Compl						
Gender=Fem	✓		✓	✓	✓	
Gender=Masc	✓		✓	✓	✓	
Gender=Neut			✓	✓		
Gender[psor]=Fem						
Gender[psor]=Masc						
Gender[psor]=Neut						
HebBinyan=HIFIL						
HebBinyan=HITPAEL						
HebBinyan=HUFAL						
HebBinyan=NIFAL						
HebBinyan=PAAL						
HebBinyan=PIEL						
HebBinyan=PUAL						
HebExistential=True						

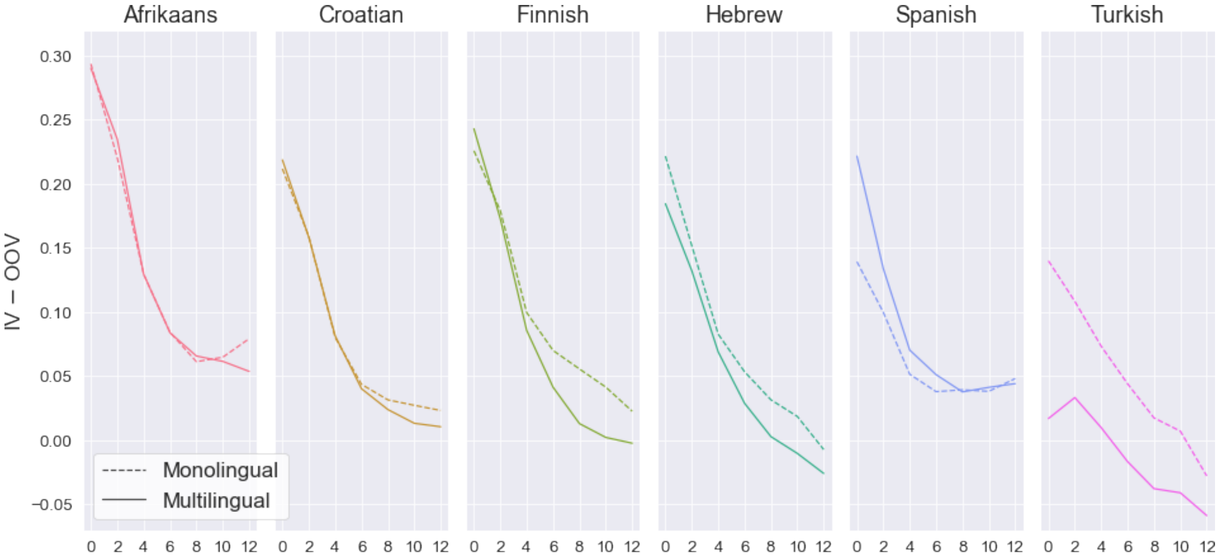
Continuation of Table B2:

Feature Labels	Arabic	Chinese	Marathi	Slovenian	Tagalog	Yorùbá
InfForm=1						
InfForm=2						
InfForm=3						
Mood=Cnd				✓		
Mood=Des						
Mood=Gen						
Mood=Imp			✓	✓		
Mood=Ind	✓		✓	✓	✓	
Mood=Nec						
Mood=Opt						
Mood=Pot						
Mood=Sub						
NumType=Card						
NumType=Dist						
NumType=Frac						
NumType=Mult						
NumType=Ord						
Number=Dual						
Number=Plur	✓	✓	✓	✓	✓	✓
Number=Sing	✓		✓	✓	✓	✓
Number[psor]=Plur						
Number[psor]=Sing						
PartForm=Agt						
PartForm=Neg						
PartForm=Past						
PartForm=Pres						
PartType=Gen						
PartType=Inf						
PartType=Neg						
Person=0						
Person=1	✓	✓	✓	✓	✓	✓
Person=2	✓	✓	✓	✓		✓
Person=3	✓	✓	✓	✓	✓	✓
Person[psor]=1						
Person[psor]=2						
Person[psor]=3						
Polarity=Neg	✓	✓	✓	✓	✓	
Polarity=Pos						
Polite=Form						
Polite=Infm						
Poss=Yes						
Prefix=Yes						
PrepCase=Npr						
PrepCase=Pre						
PronType=Art						
PronType=Dem	✓		✓	✓	✓	✓
PronType=Emp						

Continuation of Table B2:

Feature Labels	Arabic	Chinese	Marathi	Slovenian	Tagalog	Yorùbá
PronType=Ind				✓		✓
PronType=Int			✓	✓		✓
PronType=Neg				✓		
PronType=Prs	✓		✓	✓	✓	✓
PronType=Rcp						
PronType=Rel	✓		✓	✓		✓
PronType=Tot				✓		
Reflex=Yes						
Subcat=Intr						
Subcat=Prep						
Subcat=Tran						
Tense=Fut			✓	✓		
Tense=Imp						
Tense=Past			✓			
Tense=Pqp						
Tense=Pres			✓	✓		
VerbForm=Conv						
VerbForm=Fin						
VerbForm=Ger						
VerbForm=Inf						
VerbForm=Part						
VerbForm=Vnoun						
VerbType=Aux						
VerbType=Cop						
VerbType=Mod						
VerbType=Pas						
Voice=Act	✓					
Voice=Cau						
Voice=Mid						
Voice=Pass	✓	✓				

Figure D1: Generalizability of the monolingual and multilingual probes. The x-axes indicate the mBERT layer. Negative IV-OOV scores indicate instances where the probes performed *better* on OOV tokens than IV tokens.



Appendix D Monolingual probing with a dash of memorization

In §6.2, we suggest that multilingual probing is inherently more complex than monolingual probing, and that the simpler task affords the monolingual probes more opportunity to memorize the feature labels. Here, we provide additional evidence supporting this analysis.

Out-of-vocabulary words It is worthwhile to note that the UD corpora assume an open vocabulary—many of the word types in the validation and test sets do not appear during training. This allows us to evaluate the effectiveness of the probes on out-of-vocabulary (OOV) words. If the probes truly extract features versus memorizing the task, we would expect them to perform similarly on in-vocabulary (IV) and OOV words. Conversely, if the monolingual probes rely more heavily on memorization, this would predict that the multilingual probes are better able to generalize to new data.

This prediction is largely validated by the OOV tokens: We micro-averaged separate F_1 scores for the words that were seen during training and those that weren't. Since the intuition is that a probe that generalizes better will exhibit smaller gaps in performance between OOV and IV words, we subtracted the OOV scores from the IV scores to quantify how well the probes generalized to unseen words (Figure D1). For Croatian, Finnish, Hebrew, and Turkish, we observed that the gaps between IV and OOV performance tended to be smaller for the multilingual probes than the monolingual ones, especially in later layers.[8] These generalization trends suggest that the monolingual probes are more inclined towards memorization than the multilingual probes.

Language-specific features Another piece of evidence comes from language-specific features. In the multilingual experiments, we included two sets of language-specific features: Finnish infinitive forms and Hebrew verb classes (a.k.a. *binyanim*). While the monolingual probes generally outperformed their multilingual counterparts at the feature level, the opposite tended to be true for language-specific features (see Appendix E). If the multilingual probes are more extractive, especially with cross-linguistic features, this might leave the probe with more "room" to capture language-specific features (whether through extraction or memorization).

Probe complexity Given the challenges posed by doing multilabel morphosyntactic tagging in a multilingual fashion, one possibility is that a linear

[8] In contrast, for Spanish, the multilingual probes generally exhibited greater IV-OOV gaps than the monolingual models, though this trend diminished with the number of layers. Likewise, for Afrikaans, the IV-OOV gaps were very similar between the monolingual and multilingual probes. Crucially, relative to the other languages, the IV-OOV gaps were greatest for Spanish and Afrikaans (where IV performance was better) in both the monolingual and multilingual settings. This reversal of trends is likely due to their substantially larger training sets: The increased number of training tokens (and training steps) may have lured the multilingual probes to memorize the word-to-label mappings for these languages.

Table D1: Micro-averaged F_1 scores from the linear monolingual and multilingual probes (*Mono.* & *Multi.*) and the multilingual MLP-1 probes with $h = \{16, 32, 64, 128\}$ hidden dimensions.

	Mono.	*Multi.*	$h = 16$	$h = 32$	$h = 64$	$h = 128$
Afrikaans	**0.95**	0.91	0.89	0.91	0.93	0.94
Croatian	**0.92**	0.87	0.83	0.88	0.90	0.91
Finnish	**0.87**	0.83	0.77	0.83	0.85	**0.87**
Hebrew	**0.87**	0.84	0.81	0.84	0.86	**0.87**
Spanish	**0.97**	0.93	0.91	0.94	0.95	0.96
Turkish	**0.83**	0.76	0.71	0.77	0.80	0.82

Table D2: Selectivity scores from the linear monolingual and multilingual probes (*Mono.* & *Multi.*) and the multilingual MLP-1 probes with $h = \{16, 32, 64, 128\}$ hidden dimensions.

	Mono.	*Multi.*	$h = 16$	$h = 32$	$h = 64$	$h = 128$
Afrikaans	0.29	**0.50**	0.37	0.29	0.27	0.27
Croatian	0.42	**0.58**	0.42	0.39	0.39	0.39
Finnish	0.46	**0.60**	0.51	0.50	0.50	0.50
Hebrew	0.49	**0.58**	0.52	0.50	0.49	0.48
Spanish	0.35	**0.50**	0.35	0.31	0.30	0.30
Turkish	0.46	**0.47**	0.39	0.38	0.39	0.40

probe is simply not complex enough to accommodate the multilingual task. If true, this might offer an alternative explanation as to why the monolingual probes outperformed the multilingual probes.

In a small post-hoc analysis with mBERT-6, we trained multilayer perceptrons with a single hidden layer (MLP-1s) to perform the multilingual morphosyntactic tagging task. As we increased the dimensionality of the hidden layer, we found that the micro-averaged F_1 performance would approach that of the monolingual probes, but with comparable or worse selectivity. In contrast, the linear multilingual probes consistently exhibited the best selectivity. Tables D1 and D2 convey these results. In sum, these findings suggest that the improvements observed by the more complex probes resulted from them having an increased capacity for memorizing the task, rather than from being more expressive (cf. Hewitt and Liang, 2019). Thus, the advantage of the monolingual probes over the multilingual probes cannot be reduced to a linear layer not being sufficient enough to extract features from multiple languages.

Appendix E Monolingual + multilingual feature-level performance

Figures E1 though E7 report the global and feature-level F_1 results for the monolingual and multilingual probes. In the monolingual experiments, we trained separate probes for Afrikaans, Croatian, Finnish, Hebrew, Korean, Spanish, and Turkish. In a set of multilingual experiments, we then trained probes on a shuffled combination of the training data from the monolingual probes. However, we excluded the Korean dataset from these experiments, due to the lack of documentation on its construction.

Appendix F Crosslingual performance

Figure F1 shows the global F_1 results from evaluating the monolingual and multilingual probes on the held-out languages (plus Korean), while Figure F2 shows the feature-level F_1 results from evaluating the mBERT-6 probes on the held-out languages.

Figure E1: Akrikaans F$_1$

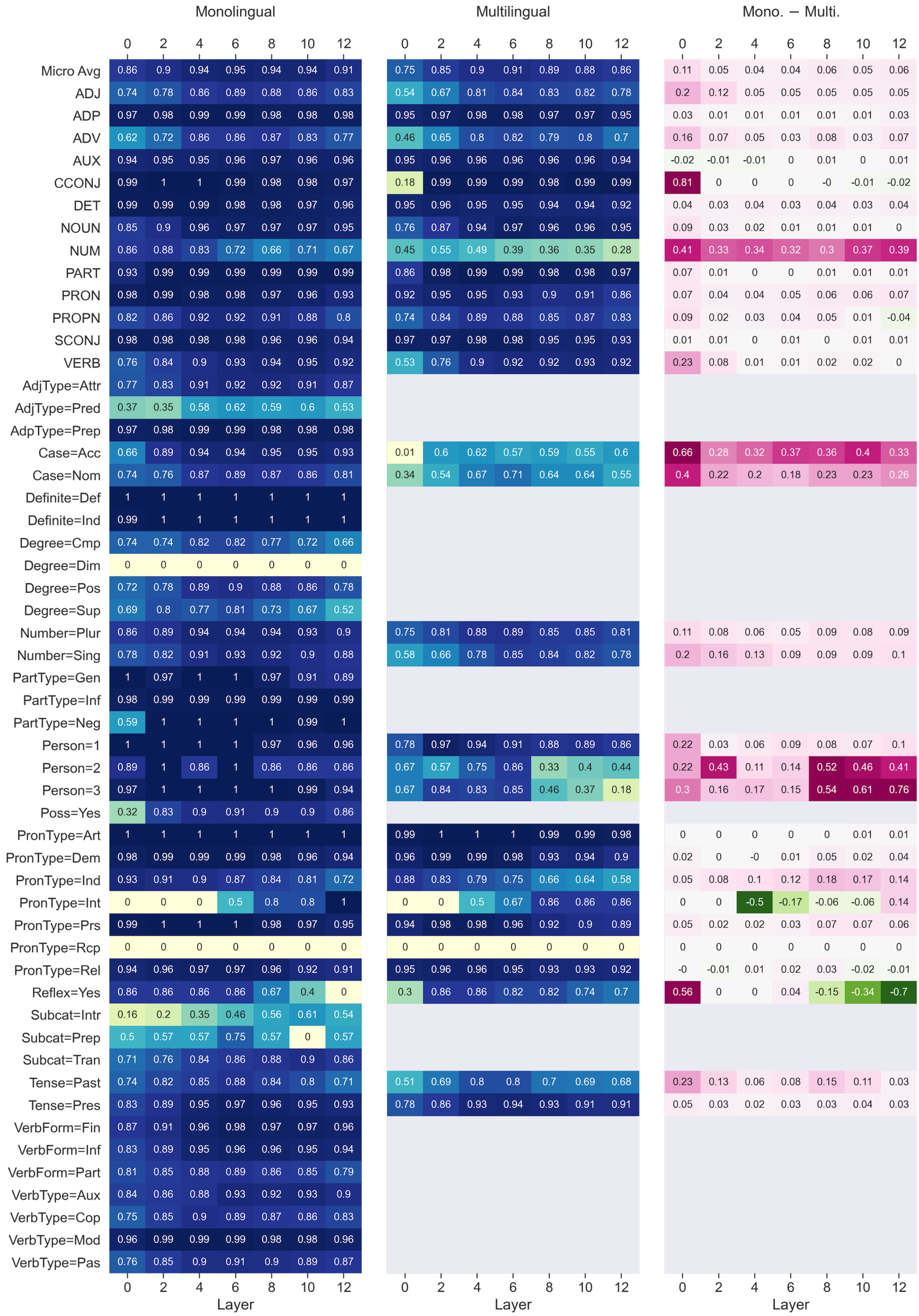
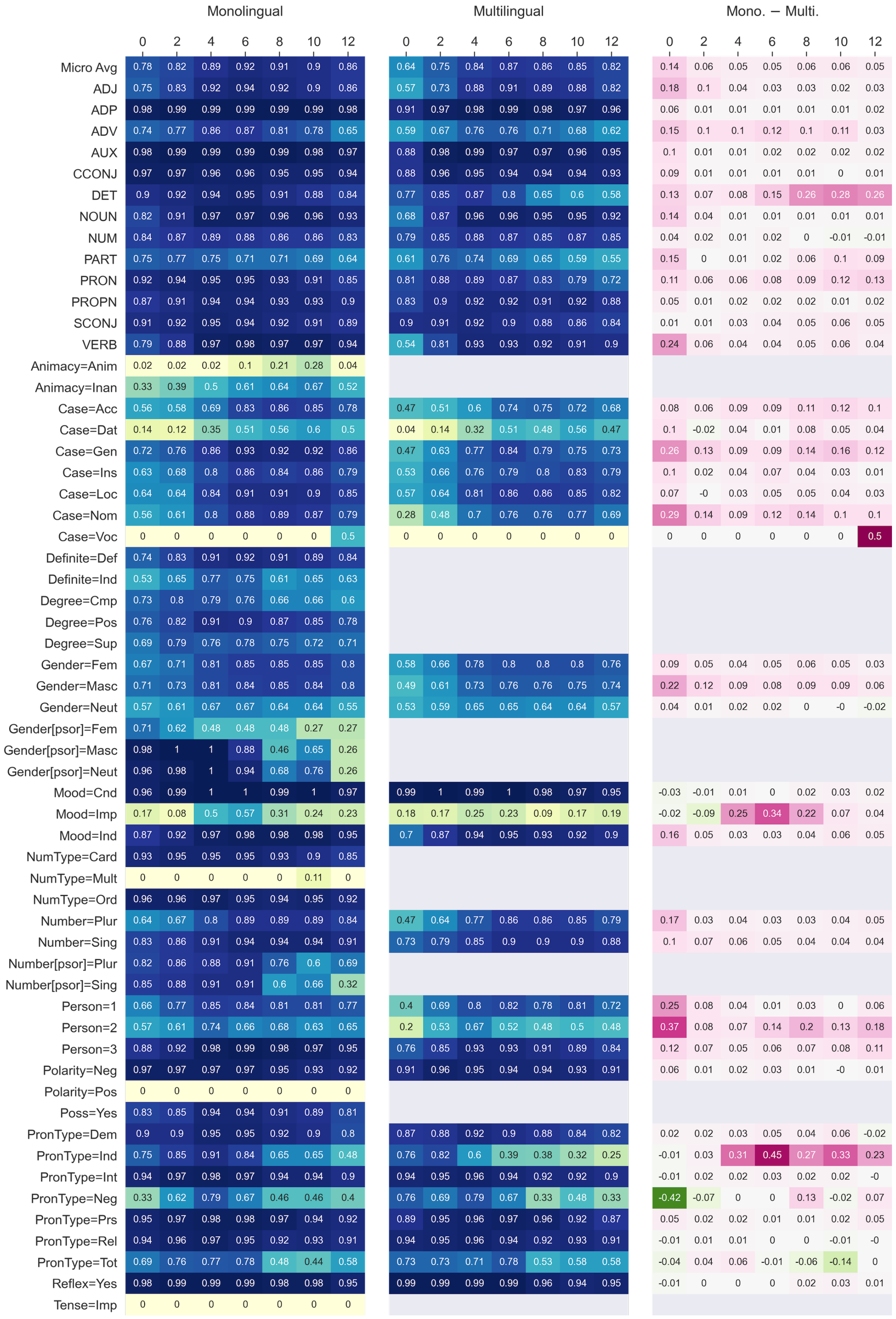


Figure E2: Croatian F$_1$



	Monolingual							Multilingual							Mono. − Multi.						
	0	2	4	6	8	10	12	0	2	4	6	8	10	12	0	2	4	6	8	10	12
Micro Avg	0.78	0.82	0.89	0.92	0.91	0.9	0.86	0.64	0.75	0.84	0.87	0.86	0.85	0.82	0.14	0.06	0.05	0.05	0.06	0.06	0.05
ADJ	0.75	0.83	0.92	0.94	0.92	0.9	0.86	0.57	0.73	0.88	0.91	0.89	0.88	0.82	0.18	0.1	0.04	0.03	0.03	0.02	0.03
ADP	0.98	0.99	0.99	0.99	0.99	0.99	0.98	0.91	0.97	0.98	0.99	0.98	0.97	0.96	0.06	0.01	0.01	0.01	0.01	0.01	0.02
ADV	0.74	0.77	0.86	0.87	0.81	0.78	0.65	0.59	0.67	0.76	0.76	0.71	0.68	0.62	0.15	0.1	0.1	0.12	0.1	0.11	0.03
AUX	0.98	0.99	0.99	0.99	0.99	0.98	0.97	0.88	0.98	0.99	0.97	0.97	0.96	0.95	0.1	0.01	0.01	0.02	0.02	0.02	0.02
CCONJ	0.97	0.97	0.96	0.96	0.95	0.95	0.94	0.88	0.96	0.95	0.94	0.94	0.94	0.93	0.09	0.01	0.01	0.01	0.01	0	0.01
DET	0.9	0.92	0.94	0.95	0.91	0.88	0.84	0.77	0.85	0.87	0.8	0.65	0.6	0.58	0.13	0.07	0.08	0.15	0.26	0.28	0.26
NOUN	0.82	0.91	0.97	0.97	0.96	0.96	0.93	0.68	0.87	0.96	0.96	0.95	0.95	0.92	0.14	0.04	0.01	0.01	0.01	0.01	0.01
NUM	0.84	0.87	0.89	0.88	0.86	0.86	0.83	0.79	0.85	0.88	0.87	0.85	0.87	0.85	0.04	0.02	0.01	0.02	0	-0.01	-0.01
PART	0.75	0.77	0.75	0.71	0.71	0.69	0.64	0.61	0.76	0.74	0.69	0.65	0.59	0.55	0.15	0	0.01	0.02	0.06	0.1	0.09
PRON	0.92	0.94	0.95	0.95	0.93	0.91	0.85	0.81	0.88	0.89	0.87	0.83	0.79	0.72	0.11	0.06	0.06	0.08	0.09	0.12	0.13
PROPN	0.87	0.91	0.94	0.94	0.93	0.93	0.9	0.83	0.9	0.92	0.92	0.91	0.92	0.88	0.05	0.01	0.02	0.02	0.02	0.01	0.02
SCONJ	0.91	0.92	0.95	0.94	0.92	0.91	0.89	0.9	0.91	0.92	0.9	0.88	0.86	0.84	0.01	0.01	0.03	0.04	0.05	0.06	0.05
VERB	0.79	0.88	0.97	0.98	0.97	0.97	0.94	0.54	0.81	0.93	0.93	0.92	0.91	0.9	0.24	0.06	0.04	0.04	0.05	0.06	0.04
Animacy=Anim	0.02	0.02	0.02	0.1	0.21	0.28	0.04														
Animacy=Inan	0.33	0.39	0.5	0.61	0.64	0.67	0.52														
Case=Acc	0.56	0.58	0.69	0.83	0.86	0.85	0.78	0.47	0.51	0.6	0.74	0.75	0.72	0.68	0.08	0.06	0.09	0.09	0.11	0.12	0.1
Case=Dat	0.14	0.12	0.35	0.51	0.56	0.6	0.5	0.04	0.14	0.32	0.51	0.48	0.56	0.47	0.1	-0.02	0.04	0.01	0.08	0.05	0.04
Case=Gen	0.72	0.76	0.86	0.93	0.92	0.92	0.86	0.47	0.63	0.77	0.84	0.79	0.75	0.73	0.26	0.13	0.09	0.09	0.14	0.16	0.12
Case=Ins	0.63	0.68	0.8	0.86	0.84	0.86	0.79	0.53	0.66	0.76	0.79	0.8	0.83	0.79	0.1	0.02	0.04	0.07	0.04	0.03	0.01
Case=Loc	0.64	0.64	0.84	0.91	0.91	0.9	0.85	0.57	0.64	0.81	0.86	0.86	0.85	0.82	0.07	-0	0.03	0.05	0.05	0.04	0.03
Case=Nom	0.56	0.61	0.8	0.88	0.89	0.87	0.79	0.28	0.48	0.7	0.76	0.76	0.77	0.69	0.29	0.14	0.09	0.12	0.14	0.1	0.1
Case=Voc	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5
Definite=Def	0.74	0.83	0.91	0.92	0.91	0.89	0.84														
Definite=Ind	0.53	0.65	0.77	0.75	0.61	0.65	0.63														
Degree=Cmp	0.73	0.8	0.79	0.76	0.66	0.66	0.6														
Degree=Pos	0.76	0.82	0.91	0.9	0.87	0.85	0.78														
Degree=Sup	0.69	0.79	0.76	0.78	0.75	0.72	0.71														
Gender=Fem	0.67	0.71	0.81	0.85	0.85	0.85	0.8	0.58	0.66	0.78	0.8	0.8	0.8	0.76	0.09	0.05	0.04	0.05	0.06	0.05	0.03
Gender=Masc	0.71	0.73	0.81	0.84	0.85	0.84	0.8	0.49	0.61	0.73	0.76	0.76	0.75	0.74	0.22	0.12	0.09	0.08	0.09	0.09	0.06
Gender=Neut	0.57	0.61	0.67	0.67	0.64	0.64	0.55	0.53	0.59	0.65	0.65	0.64	0.64	0.57	0.04	0.01	0.02	0.02	0	-0	-0.02
Gender[psor]=Fem	0.71	0.62	0.48	0.48	0.48	0.27	0.27														
Gender[psor]=Masc	0.98	1	1	0.88	0.46	0.65	0.26														
Gender[psor]=Neut	0.96	0.98	1	0.94	0.68	0.76	0.26														
Mood=Cnd	0.96	0.99	1	1	0.99	1	0.97	0.99	1	0.99	1	0.98	0.97	0.95	-0.03	-0.01	0.01	0	0.02	0.03	0.02
Mood=Imp	0.17	0.08	0.5	0.57	0.31	0.24	0.23	0.18	0.17	0.25	0.23	0.09	0.17	0.19	-0.02	-0.09	0.25	0.34	0.22	0.07	0.04
Mood=Ind	0.87	0.92	0.97	0.98	0.98	0.98	0.95	0.7	0.87	0.94	0.95	0.93	0.92	0.9	0.16	0.05	0.03	0.03	0.04	0.06	0.05
NumType=Card	0.93	0.95	0.95	0.95	0.93	0.9	0.85														
NumType=Mult	0	0	0	0	0	0.11	0														
NumType=Ord	0.96	0.96	0.97	0.95	0.94	0.95	0.92														
Number=Plur	0.64	0.67	0.8	0.89	0.89	0.89	0.84	0.47	0.64	0.77	0.86	0.86	0.85	0.79	0.17	0.03	0.04	0.03	0.03	0.04	0.05
Number=Sing	0.83	0.86	0.91	0.94	0.94	0.94	0.91	0.73	0.79	0.85	0.9	0.9	0.9	0.88	0.1	0.07	0.06	0.05	0.04	0.04	0.04
Number[psor]=Plur	0.82	0.86	0.88	0.91	0.76	0.6	0.69														
Number[psor]=Sing	0.85	0.88	0.91	0.91	0.6	0.66	0.32														
Person=1	0.66	0.77	0.85	0.84	0.81	0.81	0.77	0.4	0.69	0.8	0.82	0.78	0.81	0.72	0.25	0.08	0.04	0.01	0.03	0	0.06
Person=2	0.57	0.61	0.74	0.66	0.68	0.63	0.65	0.2	0.53	0.67	0.52	0.48	0.5	0.48	0.37	0.08	0.07	0.14	0.2	0.13	0.18
Person=3	0.88	0.92	0.98	0.99	0.98	0.97	0.95	0.76	0.85	0.93	0.93	0.91	0.89	0.84	0.12	0.07	0.05	0.06	0.07	0.08	0.11
Polarity=Neg	0.97	0.97	0.97	0.97	0.95	0.93	0.92	0.91	0.96	0.95	0.94	0.94	0.93	0.91	0.06	0.01	0.02	0.03	0.01	-0	0.01
Polarity=Pos	0	0	0	0	0	0	0														
Poss=Yes	0.83	0.85	0.94	0.94	0.91	0.89	0.81														
PronType=Dem	0.9	0.9	0.95	0.95	0.92	0.9	0.8	0.87	0.88	0.92	0.9	0.88	0.84	0.82	0.02	0.02	0.03	0.05	0.04	0.06	-0.02
PronType=Ind	0.75	0.85	0.91	0.84	0.65	0.65	0.48	0.76	0.82	0.6	0.39	0.38	0.32	0.25	-0.01	0.03	0.31	0.45	0.27	0.33	0.23
PronType=Int	0.94	0.97	0.98	0.97	0.94	0.94	0.9	0.94	0.95	0.96	0.94	0.92	0.92	0.9	-0.01	0.02	0.02	0.03	0.02	0.02	-0
PronType=Neg	0.33	0.62	0.79	0.67	0.46	0.46	0.4	0.76	0.69	0.79	0.67	0.33	0.48	0.33	-0.42	-0.07	0	0	0.13	-0.02	0.07
PronType=Prs	0.95	0.97	0.98	0.98	0.97	0.94	0.92	0.89	0.95	0.96	0.97	0.96	0.92	0.87	0.05	0.02	0.02	0.01	0.01	0.02	0.05
PronType=Rel	0.94	0.96	0.97	0.95	0.92	0.93	0.91	0.94	0.95	0.96	0.94	0.92	0.93	0.91	-0.01	0.01	0.01	0	0	-0.01	-0
PronType=Tot	0.69	0.76	0.77	0.78	0.48	0.44	0.58	0.73	0.73	0.71	0.78	0.53	0.58	0.58	-0.04	0.04	0.06	-0.01	-0.06	-0.14	0
Reflex=Yes	0.98	0.99	0.99	0.99	0.98	0.98	0.95	0.99	0.99	0.99	0.99	0.96	0.94	0.95	-0.01	0	0	0	0.02	0.03	0.01
Tense=Imp	0	0	0	0	0	0	0														

Continuation of Figure E2 (Croatian F_1):

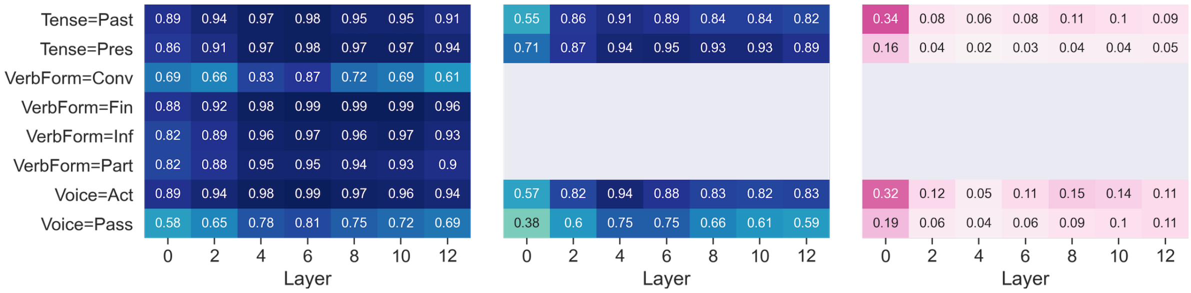
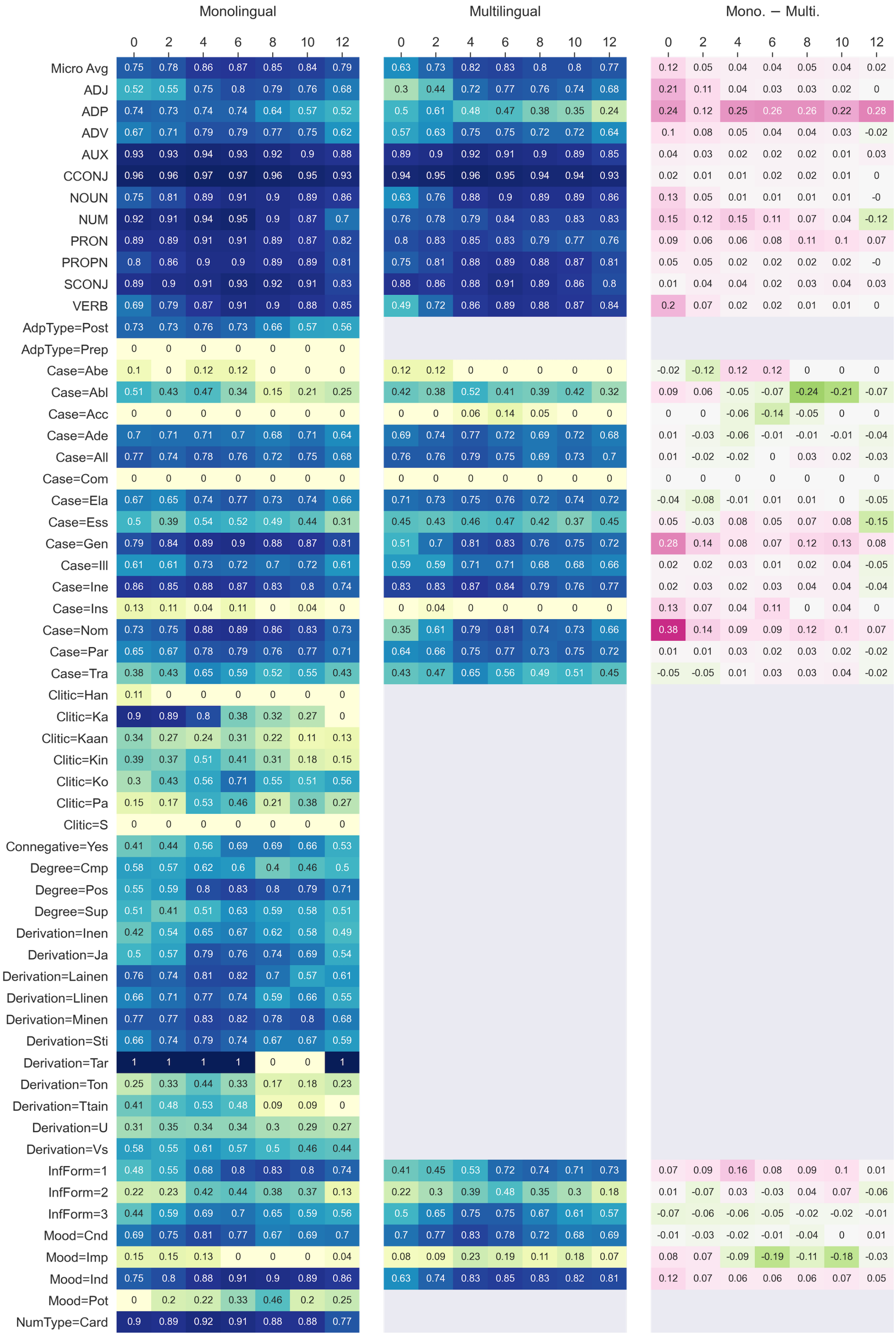
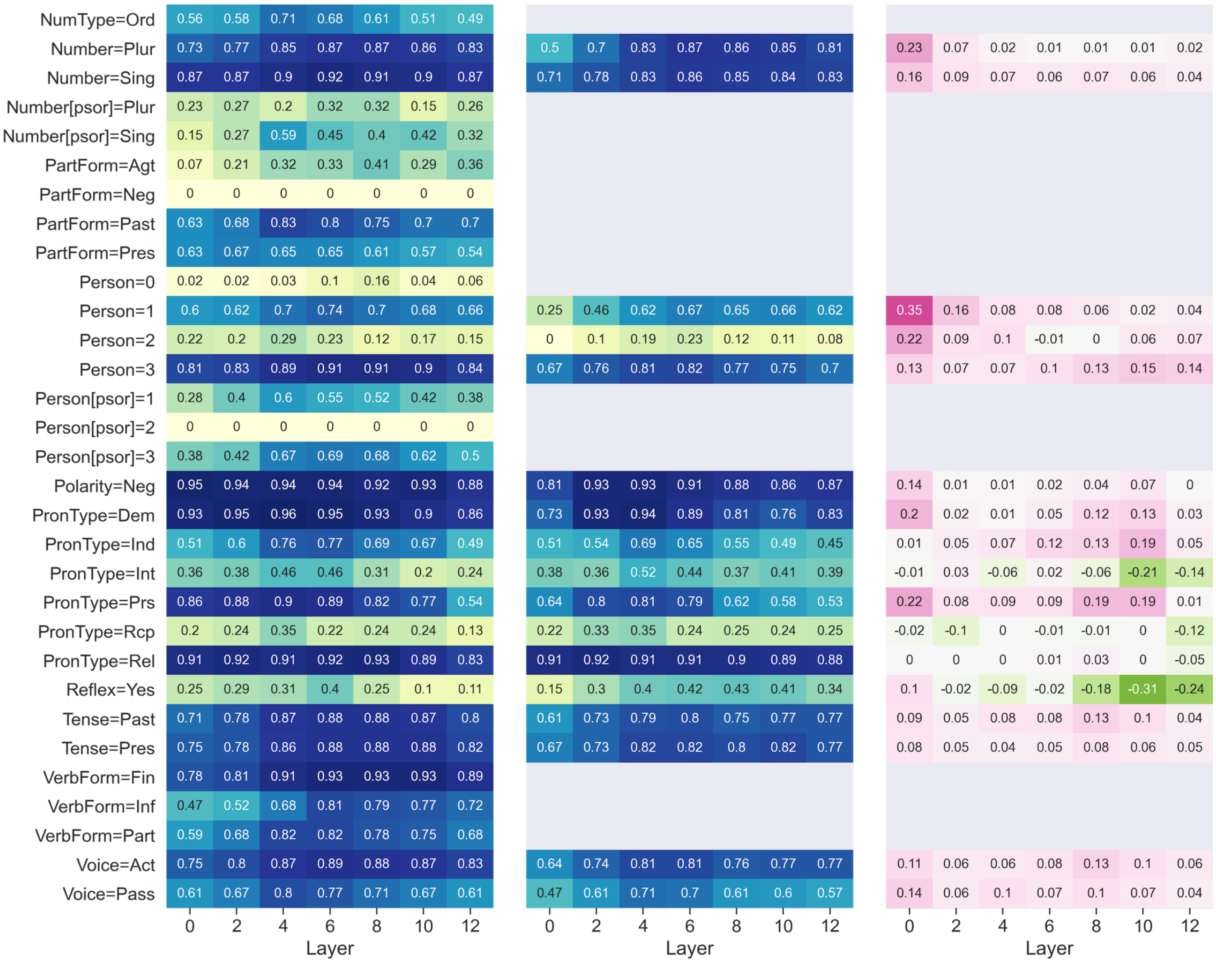


Figure E3: Finnish F_1



	Monolingual							Multilingual							Mono. − Multi.						
	0	2	4	6	8	10	12	0	2	4	6	8	10	12	0	2	4	6	8	10	12
Micro Avg	0.75	0.78	0.86	0.87	0.85	0.84	0.79	0.63	0.73	0.82	0.83	0.8	0.8	0.77	0.12	0.05	0.04	0.04	0.05	0.04	0.02
ADJ	0.52	0.55	0.75	0.8	0.79	0.76	0.68	0.3	0.44	0.72	0.77	0.76	0.74	0.68	0.21	0.11	0.04	0.03	0.03	0.02	0
ADP	0.74	0.73	0.74	0.74	0.64	0.57	0.52	0.5	0.61	0.48	0.47	0.38	0.35	0.24	0.24	0.12	0.25	0.26	0.26	0.22	0.28
ADV	0.67	0.71	0.79	0.79	0.77	0.75	0.62	0.57	0.63	0.75	0.75	0.72	0.72	0.64	0.1	0.08	0.05	0.04	0.04	0.03	-0.02
AUX	0.93	0.93	0.94	0.93	0.92	0.9	0.88	0.89	0.9	0.92	0.91	0.9	0.89	0.85	0.04	0.03	0.02	0.02	0.02	0.01	0.03
CCONJ	0.96	0.96	0.97	0.97	0.96	0.95	0.93	0.94	0.95	0.96	0.95	0.94	0.94	0.93	0.02	0.01	0.01	0.02	0.02	0.01	0
NOUN	0.75	0.81	0.89	0.91	0.9	0.89	0.86	0.63	0.76	0.88	0.9	0.89	0.89	0.86	0.13	0.05	0.01	0.01	0.01	0.01	-0
NUM	0.92	0.91	0.94	0.95	0.9	0.87	0.7	0.76	0.78	0.79	0.84	0.83	0.83	0.83	0.15	0.12	0.15	0.11	0.07	0.04	-0.12
PRON	0.89	0.89	0.91	0.91	0.89	0.87	0.82	0.8	0.83	0.85	0.83	0.79	0.77	0.76	0.09	0.06	0.06	0.08	0.11	0.1	0.07
PROPN	0.8	0.86	0.9	0.9	0.89	0.89	0.81	0.75	0.81	0.88	0.89	0.88	0.87	0.81	0.05	0.05	0.02	0.02	0.02	0.02	-0
SCONJ	0.89	0.9	0.91	0.93	0.92	0.91	0.83	0.88	0.86	0.88	0.91	0.89	0.86	0.8	0.01	0.04	0.04	0.02	0.03	0.04	0.03
VERB	0.69	0.79	0.87	0.91	0.9	0.88	0.85	0.49	0.72	0.86	0.89	0.88	0.87	0.84	0.2	0.07	0.02	0.02	0.01	0.01	0
AdpType=Post	0.73	0.73	0.76	0.73	0.66	0.57	0.56														
AdpType=Prep	0	0	0	0	0	0	0														
Case=Abe	0.1	0	0.12	0.12	0	0	0	0.12	0.12	0	0	0	0	0	-0.02	-0.12	0.12	0.12	0	0	0
Case=Abl	0.51	0.43	0.47	0.34	0.15	0.21	0.25	0.42	0.38	0.52	0.41	0.39	0.42	0.32	0.09	0.06	-0.05	-0.07	-0.24	-0.21	-0.07
Case=Acc	0	0	0	0	0	0	0	0	0	0.06	0.14	0.05	0	0	0	0	-0.06	-0.14	-0.05	0	0
Case=Ade	0.7	0.71	0.71	0.7	0.68	0.71	0.64	0.69	0.74	0.77	0.72	0.69	0.72	0.68	0.01	-0.03	-0.06	-0.01	-0.01	-0.01	-0.04
Case=All	0.77	0.74	0.78	0.76	0.72	0.75	0.68	0.76	0.76	0.79	0.75	0.69	0.73	0.7	0.01	-0.02	-0.02	0	0.03	0.02	-0.03
Case=Com	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Case=Ela	0.67	0.65	0.74	0.77	0.73	0.74	0.66	0.71	0.73	0.75	0.76	0.72	0.74	0.72	-0.04	-0.08	-0.01	0.01	0.01	0	-0.05
Case=Ess	0.5	0.39	0.54	0.52	0.49	0.44	0.31	0.45	0.43	0.46	0.47	0.42	0.37	0.45	0.05	-0.03	0.08	0.05	0.07	0.08	-0.15
Case=Gen	0.79	0.84	0.89	0.9	0.88	0.87	0.81	0.51	0.7	0.81	0.83	0.76	0.75	0.72	0.28	0.14	0.08	0.07	0.12	0.13	0.08
Case=Ill	0.61	0.61	0.73	0.72	0.7	0.72	0.61	0.59	0.59	0.71	0.71	0.68	0.68	0.66	0.02	0.02	0.03	0.01	0.02	0.04	-0.05
Case=Ine	0.86	0.85	0.88	0.87	0.83	0.8	0.74	0.83	0.83	0.87	0.84	0.79	0.76	0.77	0.02	0.03	0.02	0.03	0.04	0.04	-0.04
Case=Ins	0.13	0.11	0.04	0.11	0	0.04	0	0	0.04	0	0	0	0	0	0.13	0.07	0.04	0.11	0	0.04	0
Case=Nom	0.73	0.75	0.88	0.89	0.86	0.83	0.73	0.35	0.61	0.79	0.81	0.74	0.73	0.66	0.38	0.14	0.09	0.09	0.12	0.1	0.07
Case=Par	0.65	0.67	0.78	0.79	0.76	0.77	0.71	0.64	0.66	0.75	0.77	0.73	0.75	0.72	0.01	0.01	0.03	0.02	0.03	0.02	-0.02
Case=Tra	0.38	0.43	0.65	0.59	0.52	0.55	0.43	0.43	0.47	0.65	0.56	0.49	0.51	0.45	-0.05	-0.05	0.01	0.03	0.03	0.04	-0.02
Clitic=Han	0.11	0	0	0	0	0	0														
Clitic=Ka	0.9	0.89	0.8	0.38	0.32	0.27	0														
Clitic=Kaan	0.34	0.27	0.24	0.31	0.22	0.11	0.13														
Clitic=Kin	0.39	0.37	0.51	0.41	0.31	0.18	0.15														
Clitic=Ko	0.3	0.43	0.56	0.71	0.55	0.51	0.56														
Clitic=Pa	0.15	0.17	0.53	0.46	0.21	0.38	0.27														
Clitic=S	0	0	0	0	0	0	0														
Connegative=Yes	0.41	0.44	0.56	0.69	0.69	0.66	0.53														
Degree=Cmp	0.58	0.57	0.62	0.6	0.4	0.46	0.5														
Degree=Pos	0.55	0.59	0.8	0.83	0.8	0.79	0.71														
Degree=Sup	0.51	0.41	0.51	0.63	0.59	0.58	0.51														
Derivation=Inen	0.42	0.54	0.65	0.67	0.62	0.58	0.49														
Derivation=Ja	0.5	0.57	0.79	0.76	0.74	0.69	0.54														
Derivation=Lainen	0.76	0.74	0.81	0.82	0.7	0.57	0.61														
Derivation=Llinen	0.66	0.71	0.77	0.74	0.59	0.66	0.55														
Derivation=Minen	0.77	0.77	0.83	0.82	0.78	0.8	0.68														
Derivation=Sti	0.66	0.74	0.79	0.74	0.67	0.67	0.59														
Derivation=Tar	1	1	1	1	0	0	1														
Derivation=Ton	0.25	0.33	0.44	0.33	0.17	0.18	0.23														
Derivation=Ttain	0.41	0.48	0.53	0.48	0.09	0.09	0														
Derivation=U	0.31	0.35	0.34	0.34	0.3	0.29	0.27														
Derivation=Vs	0.58	0.55	0.61	0.57	0.5	0.46	0.44														
InfForm=1	0.48	0.55	0.68	0.8	0.83	0.8	0.74	0.41	0.45	0.53	0.72	0.74	0.71	0.73	0.07	0.09	0.16	0.08	0.09	0.1	0.01
InfForm=2	0.22	0.23	0.42	0.44	0.38	0.37	0.13	0.22	0.3	0.39	0.48	0.35	0.3	0.18	0.01	-0.07	0.03	-0.03	0.04	0.07	-0.06
InfForm=3	0.44	0.59	0.69	0.7	0.65	0.59	0.56	0.5	0.65	0.75	0.75	0.67	0.61	0.57	-0.07	-0.06	-0.06	-0.05	-0.02	-0.02	-0.01
Mood=Cnd	0.69	0.75	0.81	0.77	0.67	0.69	0.7	0.7	0.77	0.83	0.78	0.72	0.68	0.69	-0.01	-0.03	-0.02	-0.01	-0.04	0	0.01
Mood=Imp	0.15	0.15	0.13	0	0	0	0.04	0.08	0.09	0.23	0.19	0.11	0.18	0.07	0.08	0.07	-0.09	-0.19	-0.11	-0.18	-0.03
Mood=Ind	0.75	0.8	0.88	0.91	0.9	0.89	0.86	0.63	0.74	0.83	0.85	0.83	0.82	0.81	0.12	0.07	0.06	0.06	0.06	0.07	0.05
Mood=Pot	0	0.2	0.22	0.33	0.46	0.2	0.25														
NumType=Card	0.9	0.89	0.92	0.91	0.88	0.88	0.77														

Continuation of Figure E3 (Finnish F_1):



	0	2	4	6	8	10	12	0	2	4	6	8	10	12	0	2	4	6	8	10	12
NumType=Ord	0.56	0.58	0.71	0.68	0.61	0.51	0.49														
Number=Plur	0.73	0.77	0.85	0.87	0.87	0.86	0.83	0.5	0.7	0.83	0.87	0.86	0.85	0.81	0.23	0.07	0.02	0.01	0.01	0.01	0.02
Number=Sing	0.87	0.87	0.9	0.92	0.91	0.9	0.87	0.71	0.78	0.83	0.86	0.85	0.84	0.83	0.16	0.09	0.07	0.06	0.07	0.06	0.04
Number[psor]=Plur	0.23	0.27	0.2	0.32	0.32	0.15	0.26														
Number[psor]=Sing	0.15	0.27	0.59	0.45	0.4	0.42	0.32														
PartForm=Agt	0.07	0.21	0.32	0.33	0.41	0.29	0.36														
PartForm=Neg	0	0	0	0	0	0	0														
PartForm=Past	0.63	0.68	0.83	0.8	0.75	0.7	0.7														
PartForm=Pres	0.63	0.67	0.65	0.65	0.61	0.57	0.54														
Person=0	0.02	0.02	0.03	0.1	0.16	0.04	0.06														
Person=1	0.6	0.62	0.7	0.74	0.7	0.68	0.66	0.25	0.46	0.62	0.67	0.65	0.66	0.62	0.35	0.16	0.08	0.08	0.06	0.02	0.04
Person=2	0.22	0.2	0.29	0.23	0.12	0.17	0.15	0	0.1	0.19	0.23	0.12	0.11	0.08	0.22	0.09	0.1	-0.01	0	0.06	0.07
Person=3	0.81	0.83	0.89	0.91	0.91	0.9	0.84	0.67	0.76	0.81	0.82	0.77	0.75	0.7	0.13	0.07	0.07	0.1	0.13	0.15	0.14
Person[psor]=1	0.28	0.4	0.6	0.55	0.52	0.42	0.38														
Person[psor]=2	0	0	0	0	0	0	0														
Person[psor]=3	0.38	0.42	0.67	0.69	0.68	0.62	0.5														
Polarity=Neg	0.95	0.94	0.94	0.94	0.92	0.93	0.88	0.81	0.93	0.93	0.91	0.88	0.86	0.87	0.14	0.01	0.01	0.02	0.04	0.07	0
PronType=Dem	0.93	0.95	0.96	0.95	0.93	0.9	0.86	0.73	0.93	0.94	0.89	0.81	0.76	0.83	0.2	0.02	0.01	0.05	0.12	0.13	0.03
PronType=Ind	0.51	0.6	0.76	0.77	0.69	0.67	0.49	0.51	0.54	0.69	0.65	0.55	0.49	0.45	0.01	0.05	0.07	0.12	0.13	0.19	0.05
PronType=Int	0.36	0.38	0.46	0.46	0.31	0.2	0.24	0.38	0.36	0.52	0.44	0.37	0.41	0.39	-0.01	0.03	-0.06	0.02	-0.06	-0.21	-0.14
PronType=Prs	0.86	0.88	0.9	0.89	0.82	0.77	0.54	0.64	0.8	0.81	0.79	0.62	0.58	0.53	0.22	0.08	0.09	0.09	0.19	0.19	0.01
PronType=Rcp	0.2	0.24	0.35	0.22	0.24	0.24	0.13	0.22	0.33	0.35	0.24	0.25	0.24	0.25	-0.02	-0.1	0	-0.01	-0.01	0	-0.12
PronType=Rel	0.91	0.92	0.91	0.92	0.93	0.89	0.83	0.91	0.92	0.91	0.91	0.9	0.89	0.88	0	0	0	0.01	0.03	0	-0.05
Reflex=Yes	0.25	0.29	0.31	0.4	0.25	0.1	0.11	0.15	0.3	0.4	0.42	0.43	0.41	0.34	0.1	-0.02	-0.09	-0.02	-0.18	-0.31	-0.24
Tense=Past	0.71	0.78	0.87	0.88	0.88	0.87	0.8	0.61	0.73	0.79	0.8	0.75	0.77	0.77	0.09	0.05	0.08	0.08	0.13	0.1	0.04
Tense=Pres	0.75	0.78	0.86	0.88	0.88	0.88	0.82	0.67	0.73	0.82	0.82	0.8	0.82	0.77	0.08	0.05	0.04	0.05	0.08	0.06	0.05
VerbForm=Fin	0.78	0.81	0.91	0.93	0.93	0.93	0.89														
VerbForm=Inf	0.47	0.52	0.68	0.81	0.79	0.77	0.72														
VerbForm=Part	0.59	0.68	0.82	0.82	0.78	0.75	0.68														
Voice=Act	0.75	0.8	0.87	0.89	0.88	0.87	0.83	0.64	0.74	0.81	0.81	0.76	0.77	0.77	0.11	0.06	0.06	0.08	0.13	0.1	0.06
Voice=Pass	0.61	0.67	0.8	0.77	0.71	0.67	0.61	0.47	0.61	0.71	0.7	0.61	0.6	0.57	0.14	0.06	0.1	0.07	0.1	0.07	0.04

Figure E4: Hebrew F_1

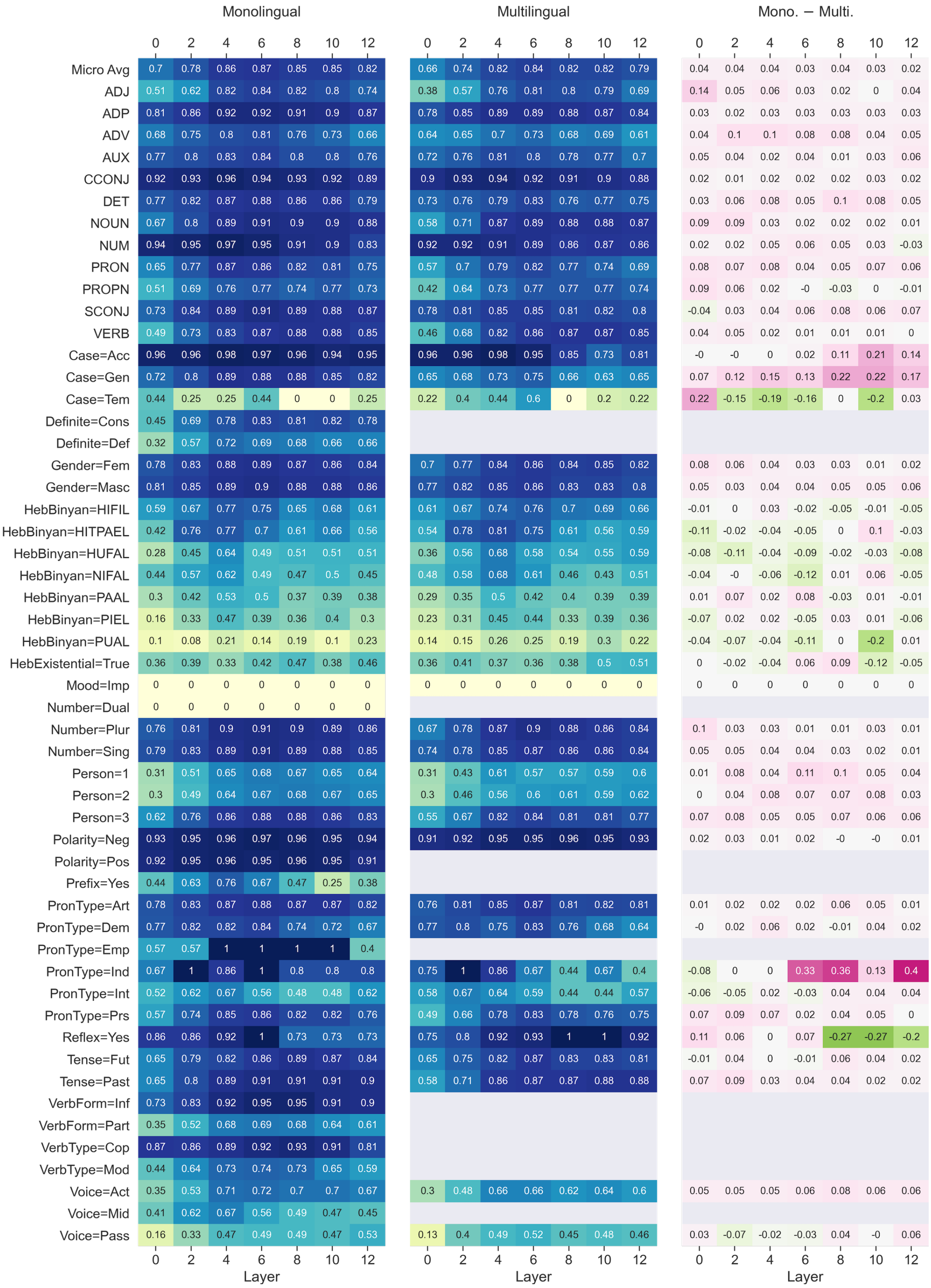
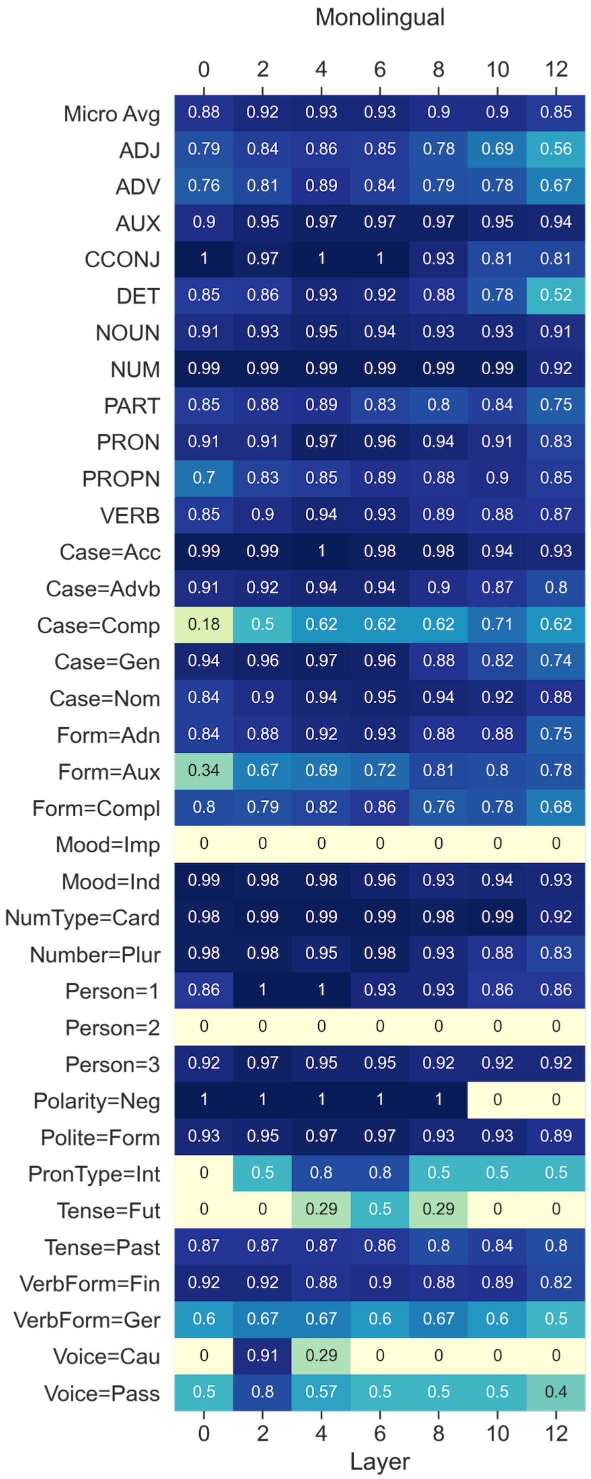
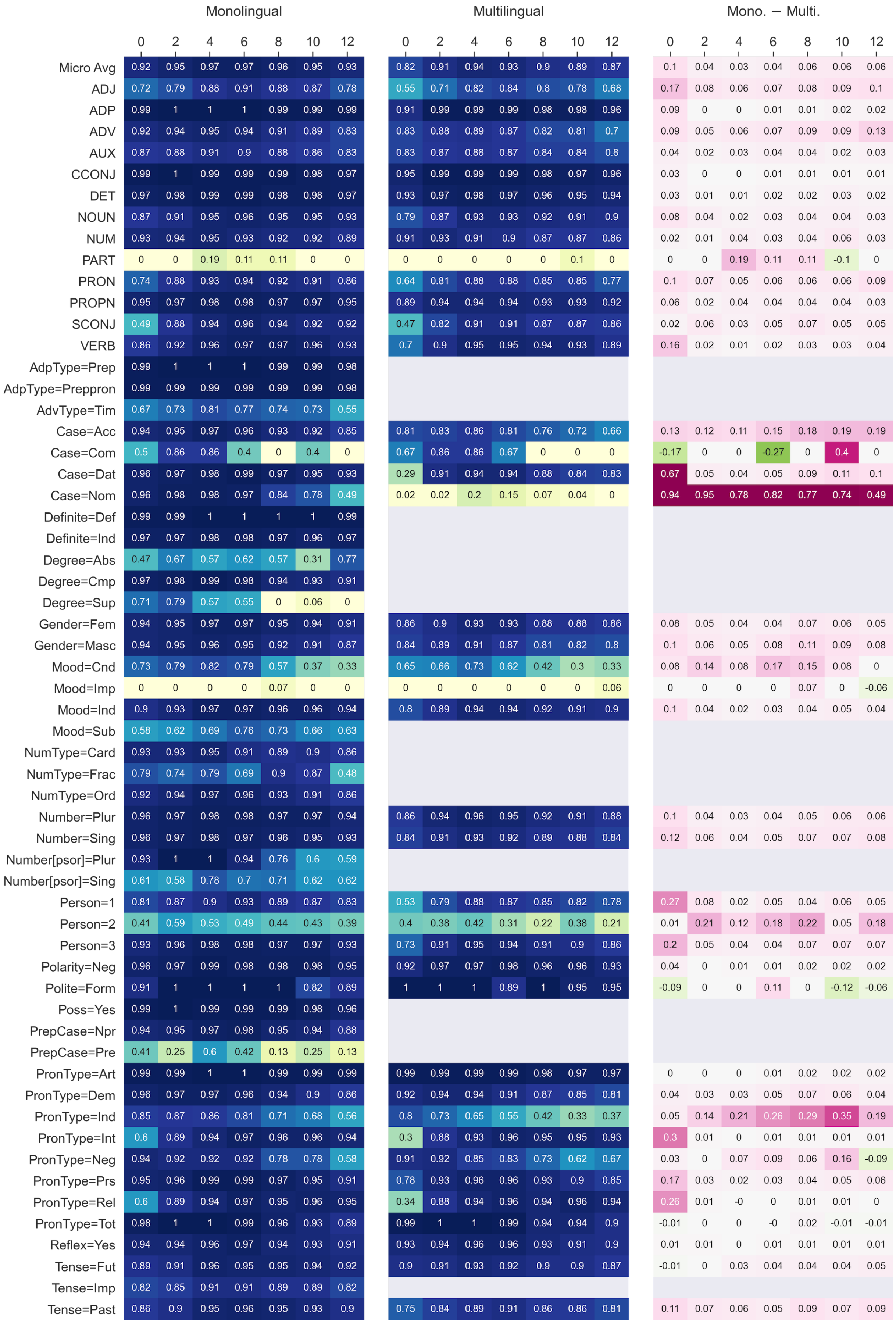


Figure E5: Korean F_1



	0	2	4	6	8	10	12
Micro Avg	0.88	0.92	0.93	0.93	0.9	0.9	0.85
ADJ	0.79	0.84	0.86	0.85	0.78	0.69	0.56
ADV	0.76	0.81	0.89	0.84	0.79	0.78	0.67
AUX	0.9	0.95	0.97	0.97	0.97	0.95	0.94
CCONJ	1	0.97	1	1	0.93	0.81	0.81
DET	0.85	0.86	0.93	0.92	0.88	0.78	0.52
NOUN	0.91	0.93	0.95	0.94	0.93	0.93	0.91
NUM	0.99	0.99	0.99	0.99	0.99	0.99	0.92
PART	0.85	0.88	0.89	0.83	0.8	0.84	0.75
PRON	0.91	0.91	0.97	0.96	0.94	0.91	0.83
PROPN	0.7	0.83	0.85	0.89	0.88	0.9	0.85
VERB	0.85	0.9	0.94	0.93	0.89	0.88	0.87
Case=Acc	0.99	0.99	1	0.98	0.98	0.94	0.93
Case=Advb	0.91	0.92	0.94	0.94	0.9	0.87	0.8
Case=Comp	0.18	0.5	0.62	0.62	0.62	0.71	0.62
Case=Gen	0.94	0.96	0.97	0.96	0.88	0.82	0.74
Case=Nom	0.84	0.9	0.94	0.95	0.94	0.92	0.88
Form=Adn	0.84	0.88	0.92	0.93	0.88	0.88	0.75
Form=Aux	0.34	0.67	0.69	0.72	0.81	0.8	0.78
Form=Compl	0.8	0.79	0.82	0.86	0.76	0.78	0.68
Mood=Imp	0	0	0	0	0	0	0
Mood=Ind	0.99	0.98	0.98	0.96	0.93	0.94	0.93
NumType=Card	0.98	0.99	0.99	0.99	0.98	0.99	0.92
Number=Plur	0.98	0.98	0.95	0.98	0.93	0.88	0.83
Person=1	0.86	1	1	0.93	0.93	0.86	0.86
Person=2	0	0	0	0	0	0	0
Person=3	0.92	0.97	0.95	0.95	0.92	0.92	0.92
Polarity=Neg	1	1	1	1	1	0	0
Polite=Form	0.93	0.95	0.97	0.97	0.93	0.93	0.89
PronType=Int	0	0.5	0.8	0.8	0.5	0.5	0.5
Tense=Fut	0	0	0.29	0.5	0.29	0	0
Tense=Past	0.87	0.87	0.87	0.86	0.8	0.84	0.8
VerbForm=Fin	0.92	0.92	0.88	0.9	0.88	0.89	0.82
VerbForm=Ger	0.6	0.67	0.67	0.6	0.67	0.6	0.5
Voice=Cau	0	0.91	0.29	0	0	0	0
Voice=Pass	0.5	0.8	0.57	0.5	0.5	0.5	0.4

Figure E6: Spanish F_1



	Monolingual							Multilingual							Mono. − Multi.						
	0	2	4	6	8	10	12	0	2	4	6	8	10	12	0	2	4	6	8	10	12
Micro Avg	0.92	0.95	0.97	0.97	0.96	0.95	0.93	0.82	0.91	0.94	0.93	0.9	0.89	0.87	0.1	0.04	0.03	0.04	0.06	0.06	0.06
ADJ	0.72	0.79	0.88	0.91	0.88	0.87	0.78	0.55	0.71	0.82	0.84	0.8	0.78	0.68	0.17	0.08	0.06	0.07	0.08	0.09	0.1
ADP	0.99	1	1	1	0.99	0.99	0.99	0.91	0.99	0.99	0.99	0.98	0.98	0.96	0.09	0	0	0.01	0.01	0.02	0.02
ADV	0.92	0.94	0.95	0.94	0.91	0.89	0.83	0.83	0.88	0.89	0.87	0.82	0.81	0.7	0.09	0.05	0.06	0.07	0.09	0.09	0.13
AUX	0.87	0.88	0.91	0.9	0.88	0.86	0.83	0.83	0.87	0.88	0.87	0.84	0.84	0.8	0.04	0.02	0.03	0.04	0.04	0.02	0.03
CCONJ	0.99	1	0.99	0.99	0.99	0.98	0.97	0.95	0.99	0.99	0.99	0.98	0.97	0.96	0.03	0	0	0.01	0.01	0.01	0.01
DET	0.97	0.98	0.99	0.99	0.98	0.98	0.97	0.93	0.97	0.98	0.97	0.96	0.95	0.94	0.03	0.01	0.01	0.02	0.02	0.03	0.02
NOUN	0.87	0.91	0.95	0.96	0.95	0.95	0.93	0.79	0.87	0.93	0.93	0.92	0.91	0.9	0.08	0.04	0.02	0.03	0.04	0.04	0.03
NUM	0.93	0.94	0.95	0.93	0.92	0.92	0.89	0.91	0.93	0.91	0.9	0.87	0.87	0.86	0.02	0.01	0.04	0.03	0.04	0.06	0.03
PART	0	0	0.19	0.11	0.11	0	0	0	0	0	0	0	0.1	0	0	0	0.19	0.11	0.11	-0.1	0
PRON	0.74	0.88	0.93	0.94	0.92	0.91	0.86	0.64	0.81	0.88	0.88	0.85	0.85	0.77	0.1	0.07	0.05	0.06	0.06	0.06	0.09
PROPN	0.95	0.97	0.98	0.98	0.97	0.97	0.95	0.89	0.94	0.94	0.94	0.93	0.93	0.92	0.06	0.02	0.04	0.04	0.04	0.04	0.03
SCONJ	0.49	0.88	0.94	0.96	0.94	0.92	0.92	0.47	0.82	0.91	0.91	0.87	0.87	0.86	0.02	0.06	0.03	0.05	0.07	0.05	0.05
VERB	0.86	0.92	0.96	0.97	0.97	0.96	0.93	0.7	0.9	0.95	0.95	0.94	0.93	0.89	0.16	0.02	0.01	0.02	0.03	0.03	0.04
AdpType=Prep	0.99	1	1	1	0.99	0.99	0.98														
AdpType=Preppron	0.99	0.99	0.99	0.99	0.99	0.99	0.98														
AdvType=Tim	0.67	0.73	0.81	0.77	0.74	0.73	0.55														
Case=Acc	0.94	0.95	0.97	0.96	0.93	0.92	0.85	0.81	0.83	0.86	0.81	0.76	0.72	0.66	0.13	0.12	0.11	0.15	0.18	0.19	0.19
Case=Com	0.5	0.86	0.86	0.4	0	0.4	0	0.67	0.86	0.86	0.67	0	0	0	-0.17	0	0	-0.27	0	0.4	0
Case=Dat	0.96	0.97	0.98	0.99	0.97	0.95	0.93	0.29	0.91	0.94	0.94	0.88	0.84	0.83	0.67	0.05	0.04	0.05	0.09	0.11	0.1
Case=Nom	0.96	0.98	0.98	0.97	0.84	0.78	0.49	0.02	0.02	0.2	0.15	0.07	0.04	0	0.94	0.95	0.78	0.82	0.77	0.74	0.49
Definite=Def	0.99	0.99	1	1	1	1	0.99														
Definite=Ind	0.97	0.97	0.98	0.98	0.97	0.96	0.97														
Degree=Abs	0.47	0.67	0.57	0.62	0.57	0.31	0.77														
Degree=Cmp	0.97	0.98	0.99	0.98	0.94	0.93	0.91														
Degree=Sup	0.71	0.79	0.57	0.55	0	0.06	0														
Gender=Fem	0.94	0.95	0.97	0.97	0.95	0.94	0.91	0.86	0.9	0.93	0.93	0.88	0.88	0.86	0.08	0.05	0.04	0.04	0.07	0.06	0.05
Gender=Masc	0.94	0.95	0.96	0.95	0.92	0.91	0.87	0.84	0.89	0.91	0.87	0.81	0.82	0.8	0.1	0.06	0.05	0.08	0.11	0.09	0.08
Mood=Cnd	0.73	0.79	0.82	0.79	0.57	0.37	0.33	0.65	0.66	0.73	0.62	0.42	0.3	0.33	0.08	0.14	0.08	0.17	0.15	0.08	0
Mood=Imp	0	0	0	0	0.07	0	0	0	0	0	0	0	0	0.06	0	0	0	0	0.07	0	-0.06
Mood=Ind	0.9	0.93	0.97	0.97	0.96	0.96	0.94	0.8	0.89	0.94	0.94	0.92	0.91	0.9	0.1	0.04	0.02	0.03	0.04	0.05	0.04
Mood=Sub	0.58	0.62	0.69	0.76	0.73	0.66	0.63														
NumType=Card	0.93	0.93	0.95	0.91	0.89	0.9	0.86														
NumType=Frac	0.79	0.74	0.79	0.69	0.9	0.87	0.48														
NumType=Ord	0.92	0.94	0.97	0.96	0.93	0.91	0.86														
Number=Plur	0.96	0.97	0.98	0.98	0.97	0.97	0.94	0.86	0.94	0.96	0.95	0.92	0.91	0.88	0.1	0.04	0.03	0.04	0.05	0.06	0.06
Number=Sing	0.96	0.97	0.98	0.97	0.96	0.95	0.93	0.84	0.91	0.93	0.92	0.89	0.88	0.84	0.12	0.06	0.04	0.05	0.07	0.07	0.08
Number[psor]=Plur	0.93	1	1	0.94	0.76	0.6	0.59														
Number[psor]=Sing	0.61	0.58	0.78	0.7	0.71	0.62	0.62														
Person=1	0.81	0.87	0.9	0.93	0.89	0.87	0.83	0.53	0.79	0.88	0.87	0.85	0.82	0.78	0.27	0.08	0.02	0.05	0.04	0.06	0.05
Person=2	0.41	0.59	0.53	0.49	0.44	0.43	0.39	0.4	0.38	0.42	0.31	0.22	0.38	0.21	0.01	0.21	0.12	0.18	0.22	0.05	0.18
Person=3	0.93	0.96	0.98	0.98	0.97	0.97	0.93	0.73	0.91	0.95	0.94	0.91	0.9	0.86	0.2	0.05	0.04	0.04	0.07	0.07	0.07
Polarity=Neg	0.96	0.97	0.99	0.98	0.98	0.98	0.95	0.92	0.97	0.97	0.98	0.96	0.96	0.93	0.04	0	0.01	0.01	0.02	0.02	0.02
Polite=Form	0.91	1	1	1	1	0.82	0.89	1	1	1	0.89	1	0.95	0.95	-0.09	0	0	0.11	0	-0.12	-0.06
Poss=Yes	0.99	1	0.99	0.99	0.99	0.98	0.96														
PrepCase=Npr	0.94	0.95	0.97	0.98	0.95	0.94	0.88														
PrepCase=Pre	0.41	0.25	0.6	0.42	0.13	0.25	0.13														
PronType=Art	0.99	0.99	1	1	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.98	0.97	0.97	0	0	0	0.01	0.02	0.02	0.02
PronType=Dem	0.96	0.97	0.97	0.96	0.94	0.9	0.86	0.92	0.94	0.94	0.91	0.87	0.85	0.81	0.04	0.03	0.03	0.05	0.07	0.06	0.04
PronType=Ind	0.85	0.87	0.86	0.81	0.71	0.68	0.56	0.8	0.73	0.65	0.55	0.42	0.33	0.37	0.05	0.14	0.21	0.26	0.29	0.35	0.19
PronType=Int	0.6	0.89	0.94	0.97	0.96	0.96	0.94	0.3	0.88	0.93	0.96	0.95	0.95	0.93	0.3	0.01	0	0.01	0.01	0.01	0.01
PronType=Neg	0.94	0.92	0.92	0.92	0.78	0.78	0.58	0.91	0.92	0.85	0.83	0.73	0.62	0.67	0.03	0	0.07	0.09	0.06	0.16	-0.09
PronType=Prs	0.95	0.96	0.99	0.99	0.97	0.95	0.91	0.78	0.93	0.96	0.96	0.93	0.9	0.85	0.17	0.03	0.02	0.03	0.04	0.05	0.06
PronType=Rel	0.6	0.89	0.94	0.97	0.95	0.96	0.95	0.34	0.88	0.94	0.96	0.94	0.96	0.94	0.26	0.01	-0	0	0.01	0.01	0
PronType=Tot	0.98	1	1	0.99	0.96	0.93	0.89	0.99	1	1	0.99	0.94	0.94	0.9	-0.01	0	0	-0	0.02	-0.01	-0.01
Reflex=Yes	0.94	0.94	0.96	0.97	0.94	0.93	0.91	0.93	0.94	0.96	0.96	0.93	0.91	0.9	0.01	0.01	0	0.01	0.01	0.01	0.01
Tense=Fut	0.89	0.91	0.96	0.95	0.95	0.94	0.92	0.9	0.91	0.93	0.92	0.9	0.9	0.87	-0.01	0	0.03	0.04	0.04	0.04	0.05
Tense=Imp	0.82	0.85	0.91	0.91	0.89	0.89	0.82														
Tense=Past	0.86	0.9	0.95	0.96	0.95	0.93	0.9	0.75	0.84	0.89	0.91	0.86	0.86	0.81	0.11	0.07	0.06	0.05	0.09	0.07	0.09

Continuation of Figure E6 (Spanish F_1):

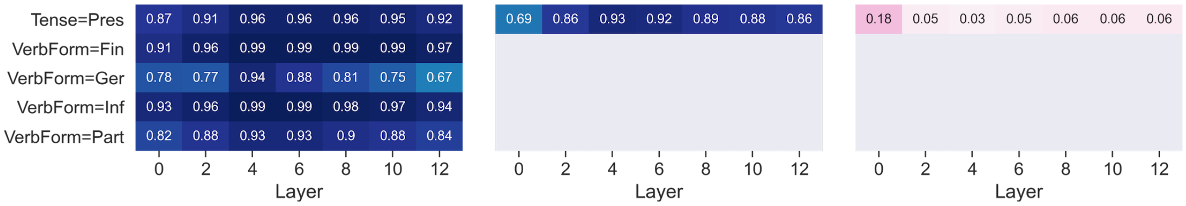
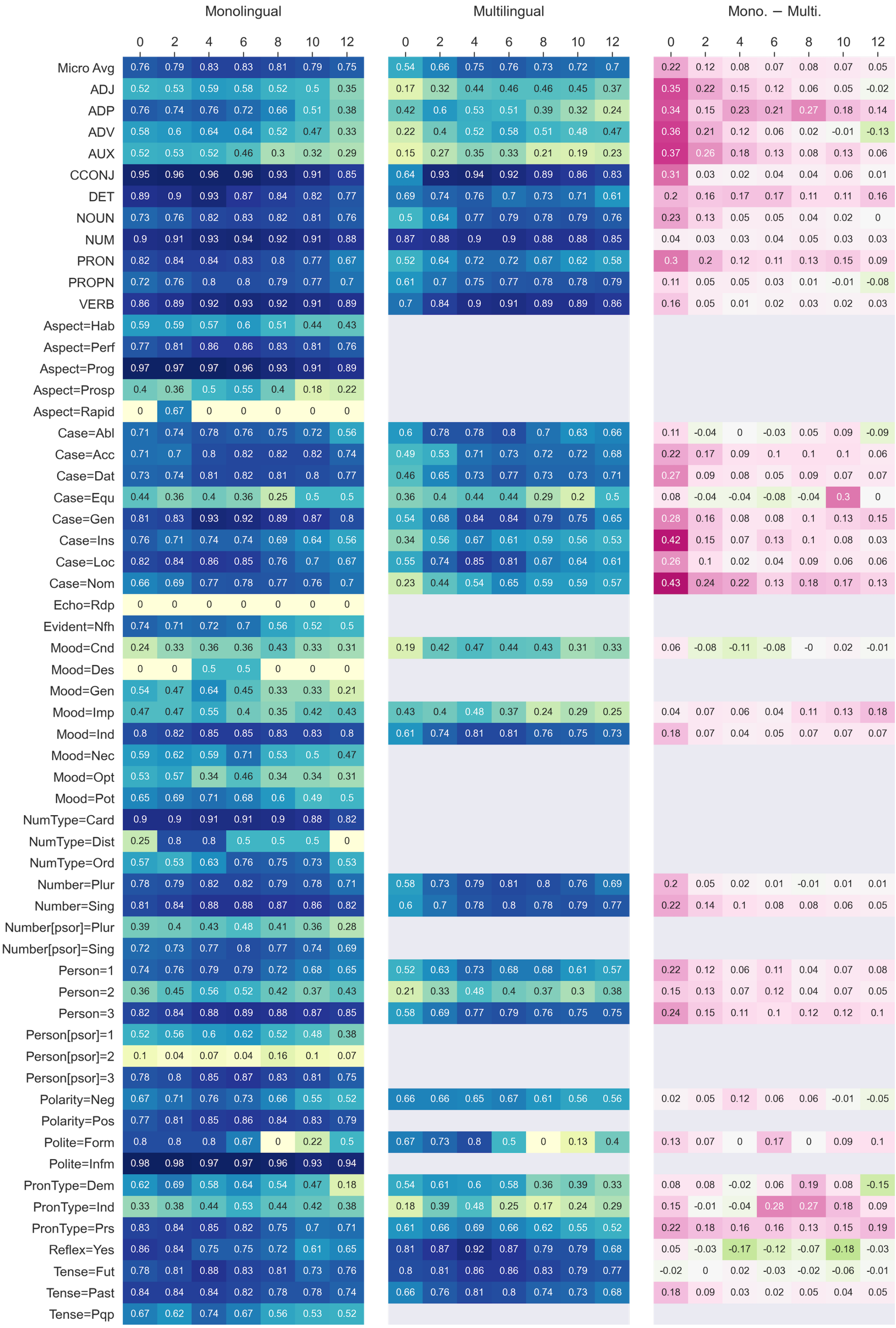


Figure E7: Turkish F_1



	Monolingual							Multilingual							Mono. − Multi.						
	0	2	4	6	8	10	12	0	2	4	6	8	10	12	0	2	4	6	8	10	12
Micro Avg	0.76	0.79	0.83	0.83	0.81	0.79	0.75	0.54	0.66	0.75	0.76	0.73	0.72	0.7	0.22	0.12	0.08	0.07	0.08	0.07	0.05
ADJ	0.52	0.53	0.59	0.58	0.52	0.5	0.35	0.17	0.32	0.44	0.46	0.46	0.45	0.37	0.35	0.22	0.15	0.12	0.06	0.05	-0.02
ADP	0.76	0.74	0.76	0.72	0.66	0.51	0.38	0.42	0.6	0.53	0.51	0.39	0.32	0.24	0.34	0.15	0.23	0.21	0.27	0.18	0.14
ADV	0.58	0.6	0.64	0.64	0.52	0.47	0.33	0.22	0.4	0.52	0.58	0.51	0.48	0.47	0.36	0.21	0.12	0.06	0.02	-0.01	-0.13
AUX	0.52	0.53	0.52	0.46	0.3	0.32	0.29	0.15	0.27	0.35	0.33	0.21	0.19	0.23	0.37	0.26	0.18	0.13	0.08	0.13	0.06
CCONJ	0.95	0.96	0.96	0.96	0.93	0.91	0.85	0.64	0.93	0.94	0.92	0.89	0.86	0.83	0.31	0.03	0.02	0.04	0.04	0.06	0.01
DET	0.89	0.9	0.93	0.87	0.84	0.82	0.77	0.69	0.74	0.76	0.7	0.73	0.71	0.61	0.2	0.16	0.17	0.17	0.11	0.11	0.16
NOUN	0.73	0.76	0.82	0.83	0.82	0.81	0.76	0.5	0.64	0.77	0.79	0.78	0.79	0.76	0.23	0.13	0.05	0.05	0.04	0.02	0
NUM	0.9	0.91	0.93	0.94	0.92	0.91	0.88	0.87	0.88	0.9	0.9	0.88	0.88	0.85	0.04	0.03	0.03	0.04	0.05	0.03	0.03
PRON	0.82	0.84	0.84	0.83	0.8	0.77	0.67	0.52	0.64	0.72	0.72	0.67	0.62	0.58	0.3	0.2	0.12	0.11	0.13	0.15	0.09
PROPN	0.72	0.76	0.8	0.8	0.79	0.77	0.7	0.61	0.7	0.75	0.77	0.78	0.78	0.79	0.11	0.05	0.05	0.03	0.01	-0.01	-0.08
VERB	0.86	0.89	0.92	0.93	0.92	0.91	0.89	0.7	0.84	0.9	0.91	0.89	0.89	0.86	0.16	0.05	0.01	0.02	0.03	0.02	0.03
Aspect=Hab	0.59	0.59	0.57	0.6	0.51	0.44	0.43														
Aspect=Perf	0.77	0.81	0.86	0.86	0.83	0.81	0.76														
Aspect=Prog	0.97	0.97	0.97	0.96	0.93	0.91	0.89														
Aspect=Prosp	0.4	0.36	0.5	0.55	0.4	0.18	0.22														
Aspect=Rapid	0	0.67	0	0	0	0	0														
Case=Abl	0.71	0.74	0.78	0.76	0.75	0.72	0.56	0.6	0.78	0.78	0.8	0.7	0.63	0.66	0.11	-0.04	0	-0.03	0.05	0.09	-0.09
Case=Acc	0.71	0.7	0.8	0.82	0.82	0.82	0.74	0.49	0.53	0.71	0.73	0.72	0.72	0.68	0.22	0.17	0.09	0.1	0.1	0.1	0.06
Case=Dat	0.73	0.74	0.81	0.82	0.81	0.8	0.77	0.46	0.65	0.73	0.77	0.73	0.73	0.71	0.27	0.09	0.08	0.05	0.09	0.07	0.07
Case=Equ	0.44	0.36	0.4	0.36	0.25	0.5	0.5	0.36	0.4	0.44	0.44	0.29	0.2	0.5	0.08	-0.04	-0.04	-0.08	-0.04	0.3	0
Case=Gen	0.81	0.83	0.93	0.92	0.89	0.87	0.8	0.54	0.68	0.84	0.84	0.79	0.75	0.65	0.28	0.16	0.08	0.08	0.1	0.13	0.15
Case=Ins	0.76	0.71	0.74	0.74	0.69	0.64	0.56	0.34	0.56	0.67	0.61	0.59	0.56	0.53	0.42	0.15	0.07	0.13	0.1	0.08	0.03
Case=Loc	0.82	0.84	0.86	0.85	0.76	0.7	0.67	0.55	0.74	0.85	0.81	0.67	0.64	0.61	0.26	0.1	0.02	0.04	0.09	0.06	0.06
Case=Nom	0.66	0.69	0.77	0.78	0.77	0.76	0.7	0.23	0.44	0.54	0.65	0.59	0.59	0.57	0.43	0.24	0.22	0.13	0.18	0.17	0.13
Echo=Rdp	0	0	0	0	0	0	0														
Evident=Nfh	0.74	0.71	0.72	0.7	0.56	0.52	0.5														
Mood=Cnd	0.24	0.33	0.36	0.36	0.43	0.33	0.31	0.19	0.42	0.47	0.44	0.43	0.31	0.33	0.06	-0.08	-0.11	-0.08	-0	0.02	-0.01
Mood=Des	0	0	0.5	0.5	0	0	0														
Mood=Gen	0.54	0.47	0.64	0.45	0.33	0.33	0.21														
Mood=Imp	0.47	0.47	0.55	0.4	0.35	0.42	0.43	0.43	0.4	0.48	0.37	0.24	0.29	0.25	0.04	0.07	0.06	0.04	0.11	0.13	0.18
Mood=Ind	0.8	0.82	0.85	0.85	0.83	0.83	0.8	0.61	0.74	0.81	0.81	0.76	0.75	0.73	0.18	0.07	0.04	0.05	0.07	0.07	0.07
Mood=Nec	0.59	0.62	0.59	0.71	0.53	0.5	0.47														
Mood=Opt	0.53	0.57	0.34	0.46	0.34	0.34	0.31														
Mood=Pot	0.65	0.69	0.71	0.68	0.6	0.49	0.5														
NumType=Card	0.9	0.9	0.91	0.91	0.9	0.88	0.82														
NumType=Dist	0.25	0.8	0.8	0.5	0.5	0.5	0														
NumType=Ord	0.57	0.53	0.63	0.76	0.75	0.73	0.53														
Number=Plur	0.78	0.79	0.82	0.82	0.79	0.78	0.71	0.58	0.73	0.79	0.81	0.8	0.76	0.69	0.2	0.05	0.02	0.01	-0.01	0.01	0.01
Number=Sing	0.81	0.84	0.88	0.88	0.87	0.86	0.82	0.6	0.7	0.78	0.8	0.78	0.79	0.77	0.22	0.14	0.1	0.08	0.08	0.06	0.05
Number[psor]=Plur	0.39	0.4	0.43	0.48	0.41	0.36	0.28														
Number[psor]=Sing	0.72	0.73	0.77	0.8	0.77	0.74	0.69														
Person=1	0.74	0.76	0.79	0.79	0.72	0.68	0.65	0.52	0.63	0.73	0.68	0.68	0.61	0.57	0.22	0.12	0.06	0.11	0.04	0.07	0.08
Person=2	0.36	0.45	0.56	0.52	0.42	0.37	0.43	0.21	0.33	0.48	0.4	0.37	0.3	0.38	0.15	0.13	0.07	0.12	0.04	0.07	0.05
Person=3	0.82	0.84	0.88	0.89	0.88	0.87	0.85	0.58	0.69	0.77	0.79	0.76	0.75	0.75	0.24	0.15	0.11	0.1	0.12	0.12	0.1
Person[psor]=1	0.52	0.56	0.6	0.62	0.52	0.48	0.38														
Person[psor]=2	0.1	0.04	0.07	0.04	0.16	0.1	0.07														
Person[psor]=3	0.78	0.8	0.85	0.87	0.83	0.81	0.75														
Polarity=Neg	0.67	0.71	0.76	0.73	0.66	0.55	0.52	0.66	0.66	0.65	0.67	0.61	0.56	0.56	0.02	0.05	0.12	0.06	0.06	-0.01	-0.05
Polarity=Pos	0.77	0.81	0.85	0.86	0.84	0.83	0.79														
Polite=Form	0.8	0.8	0.8	0.67	0	0.22	0.5	0.67	0.73	0.8	0.5	0	0.13	0.4	0.13	0.07	0	0.17	0	0.09	0.1
Polite=Infm	0.98	0.98	0.97	0.97	0.96	0.93	0.94														
PronType=Dem	0.62	0.69	0.58	0.64	0.54	0.47	0.18	0.54	0.61	0.6	0.58	0.36	0.39	0.33	0.08	0.08	-0.02	0.06	0.19	0.08	-0.15
PronType=Ind	0.33	0.38	0.44	0.53	0.44	0.42	0.38	0.18	0.39	0.48	0.25	0.17	0.24	0.29	0.15	-0.01	-0.04	0.28	0.27	0.18	0.09
PronType=Prs	0.83	0.84	0.85	0.82	0.75	0.7	0.71	0.61	0.66	0.69	0.66	0.62	0.55	0.52	0.22	0.18	0.16	0.16	0.13	0.15	0.19
Reflex=Yes	0.86	0.84	0.75	0.75	0.72	0.61	0.65	0.81	0.87	0.92	0.87	0.79	0.79	0.68	0.05	-0.03	-0.17	-0.12	-0.07	-0.18	-0.03
Tense=Fut	0.78	0.81	0.88	0.83	0.81	0.73	0.76	0.8	0.81	0.86	0.86	0.83	0.79	0.77	-0.02	0	0.02	-0.03	-0.02	-0.06	-0.01
Tense=Past	0.84	0.84	0.84	0.82	0.78	0.78	0.74	0.66	0.76	0.81	0.8	0.74	0.73	0.68	0.18	0.09	0.03	0.02	0.05	0.04	0.05
Tense=Pqp	0.67	0.62	0.74	0.67	0.56	0.53	0.52														

Continuation of Figure E7 (Turkish F_1):

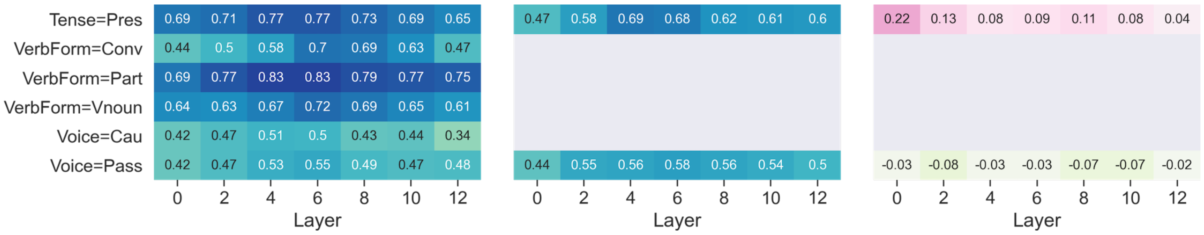


Figure F1: Micro-averaged F_1 results from evaluating the monolingual and multilingual probes on the "held-out" languages (plus Korean). The x-axes indicate the mBERT layer.

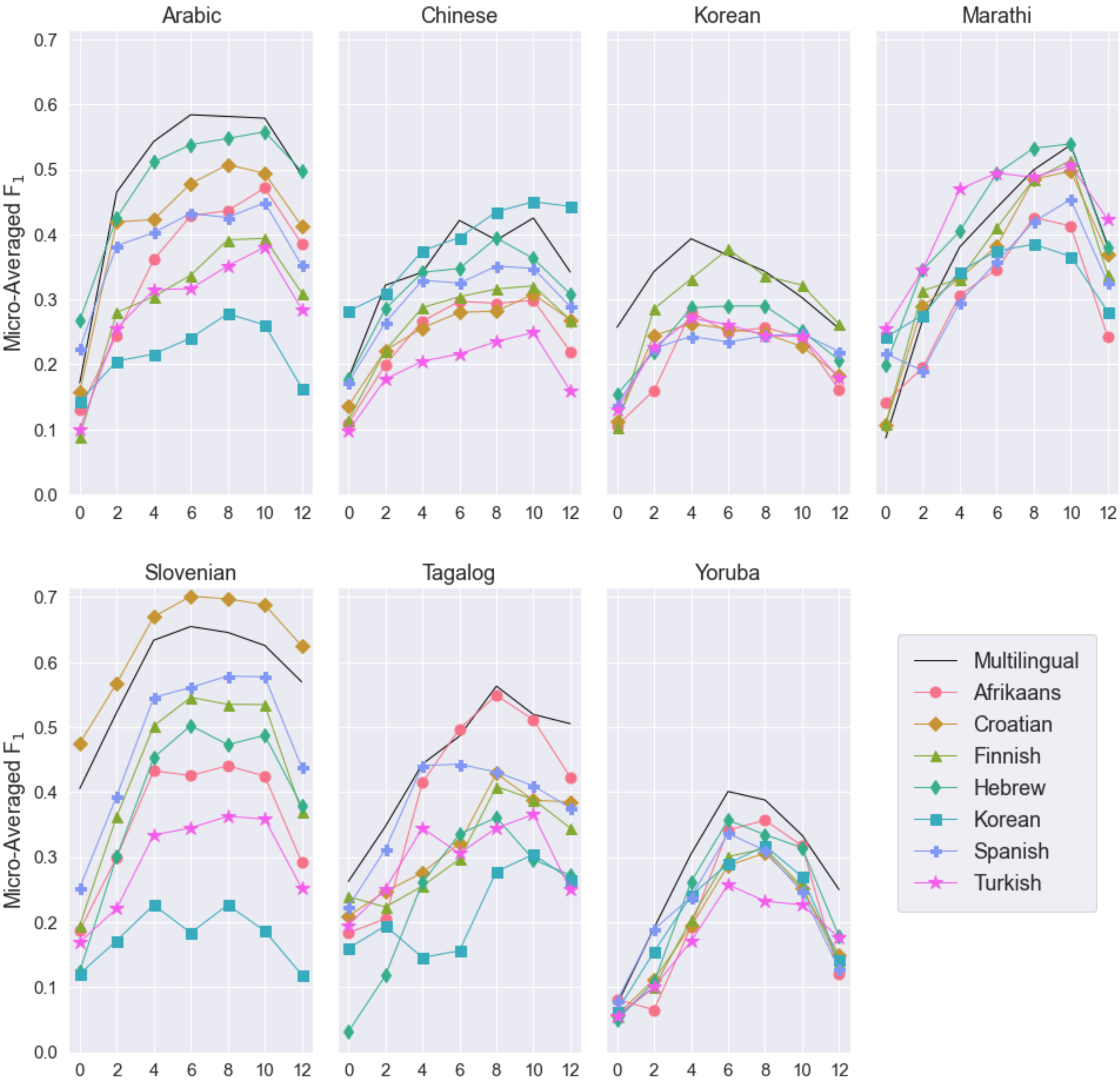
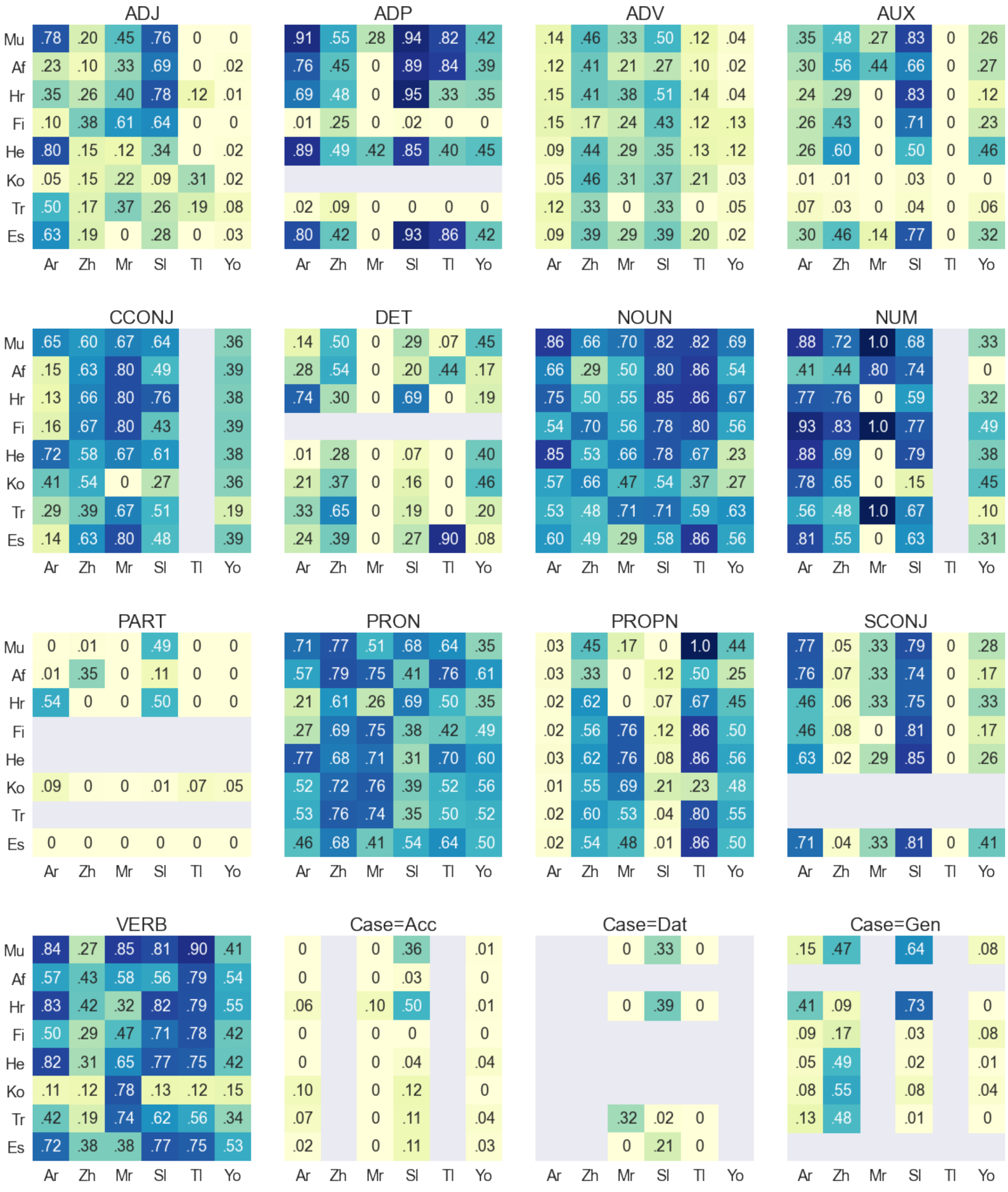
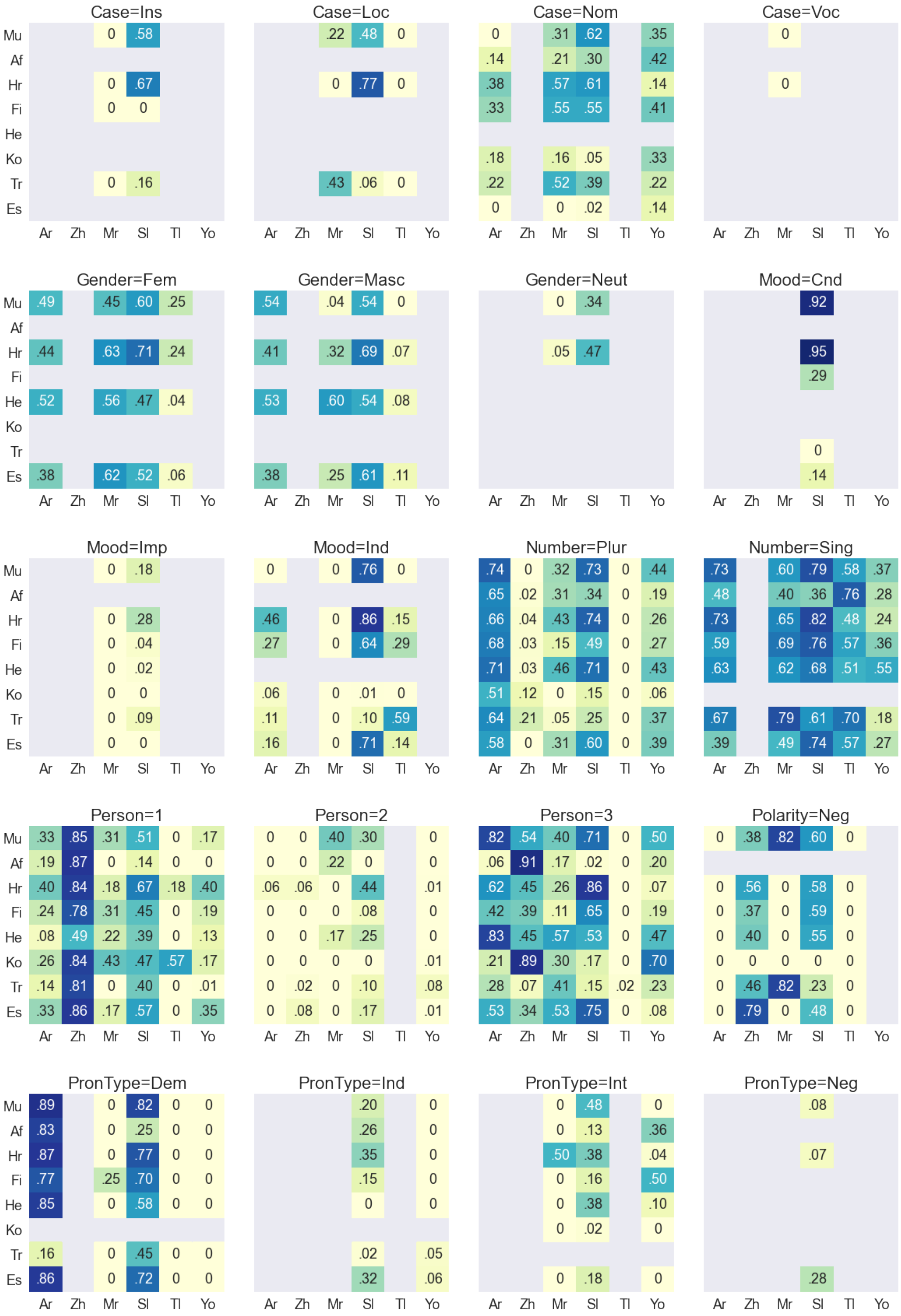


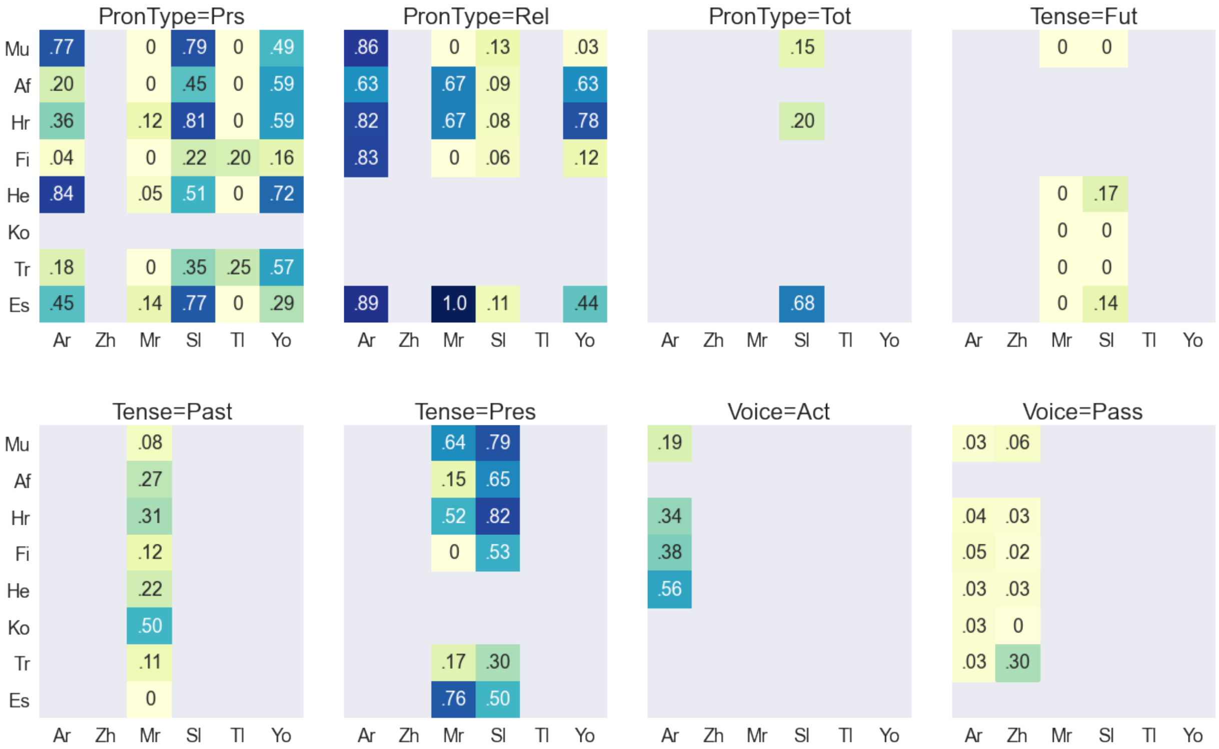
Figure F2: F_1 results from evaluating the monolingual and multilingual mBERT-6 probes on the "held-out" languages. The x-axes indicate the held-out language (Ar=Arabic, Zh=Chinese, Mr=Marathi, Sl=Slovenian, Tl=Tagalog, and Yo=Yorùbá) and the y-axes indicate the probe (Mu=Multilingual, Af=Afrikaans, Hr=Croatian, Fi=Finnish, He=Hebrew, Ko=Korean, Es=Spanish, and Tr=Turkish). Grayed-out regions indicate where the feature is not applicable to the language or annotated in the language's corpus.



Continuation of Figure F2:



Continuation of Figure F2:


PronType=Prs
Mu .77 0 .79 0 .49
Af .20 0 .45 0 .59
Hr .36 .12 .81 0 .59
Fi .04 0 .22 .20 .16
He .84 .05 .51 0 .72
Ko
Tr .18 0 .35 .25 .57
Es .45 .14 .77 0 .29
Ar Zh Mr Sl Tl Yo
PronType=Rel
Ar Zh Mr Sl Tl Yo
PronType=Tot
Ar Zh Mr Sl Tl Yo
Tense=Fut
Ar Zh Mr Sl Tl Yo
Tense=Past
Mu .08
Af .27
Hr .31
Fi .12
He .22
Ko .50
Tr .11
Es 0
Ar Zh Mr Sl Tl Yo
Tense=Pres
Ar Zh Mr Sl Tl Yo
Voice=Act
Ar Zh Mr Sl Tl Yo
Voice=Pass
Ar Zh Mr Sl Tl Yo