Analyzing the Mono- and Cross-Lingual Pretraining Dynamics of Multilingual Language Models

Terra Blevins[1] **Hila Gonen**[1,2] **Luke Zettlemoyer**[1,2]
[1] Paul G. Allen School of Computer Science & Engineering, University of Washington
[2] Meta AI Research
{blvns, lsz}@cs.washington.edu
hilagnn@gmail.com

Abstract

The emergent cross-lingual transfer seen in multilingual pretrained models has sparked significant interest in studying their behavior. However, because these analyses have focused on fully trained multilingual models, little is known about the dynamics of the multilingual pretraining process. We investigate *when* these models acquire their in-language and cross-lingual abilities by probing checkpoints taken from throughout XLM-R pretraining, using a suite of linguistic tasks. Our analysis shows that the model achieves high in-language performance early on, with lower-level linguistic skills acquired before more complex ones. In contrast, the point in pretraining when the model learns to transfer cross-lingually differs across language pairs. Interestingly, we also observe that, across many languages and tasks, the final model layer exhibits significant performance degradation over time, while linguistic knowledge propagates to lower layers of the network. Taken together, these insights highlight the complexity of multilingual pretraining and the resulting varied behavior for different languages over time.

1 Introduction

Large-scale language models pretrained jointly on text from many different languages (Delvin, 2019; Lample and Conneau, 2019; Lin et al., 2021) perform very well on various languages and on cross-lingual transfer between them (e.g., Kondratyuk and Straka, 2019; Pasini et al., 2021). Due to this success, there has been a great deal of interest in uncovering what these models learn from the multilingual pretraining signal (§6). However, these works analyze a single model artifact: the final training checkpoint at which the model is considered to be converged. Recent work has also studied monolingual models by expanding the analysis to multiple pretraining checkpoints to see how model knowledge changes across time (Liu et al., 2021).

We analyze multilingual training checkpoints throughout the pretraining process in order to identify when multilingual models obtain their in-language and cross-lingual abilities. The case of multilingual language models is particularly interesting, as the model learns both to capture individual languages and to transfer between them just from unbalanced multitask language modeling for each language.

Specifically, we retrain a popular multilingual model, XLM-R (Conneau et al., 2020a), and run a suite of linguistic tasks covering 59 languages on checkpoints from across the pretraining process.[1] This suite evaluates different syntactic and semantic skills in both monolingual and cross-lingual transfer settings. While our analysis primarily focuses on the knowledge captured in model output representations over time, we also consider how the performance of internal layers changes during pretraining for a subset of tasks.

Our analysis uncovers several insights into multilingual knowledge acquisition. First, while the model acquires most in-language linguistic information early on, cross-lingual transfer is learned across the entire pretraining process. Second, the order in which the model acquires linguistic information for each language is generally consistent with monolingual models: lower-level syntax is learned prior to higher-level syntax and then semantics. In comparison, the order in which the model learns to transfer linguistic information between specific languages can vary wildly.

Finally, we observe significant degradation of performance for many languages at the final layer of the last, converged model checkpoint. However, lower layers of the network often continue to improve later in pretraining and outperform the final layer, particularly for cross-lingual transfer. These observations indicate that there is not a single time

[1] The XLM-R$_{replica}$ checkpoints are available at https://nlp.cs.washington.edu/xlmr-across-time.

Task	Setup	Num. Langs (Pairs) In-lang.	 X-lang.	Example
BPC	Masked LM	94	–	quick The [MASK] brown fox jumps
POS Tagging	Token Labeling	44	18 → 18	ADJ The quick brown fox jumps
Dependency Arc Pred.	Token Pair Labeling	44	18 → 18	✗ ✓ The quick brown fox jumps
Dependency Arc Class.	Token Pair Labeling	44	18 → 18	amod The quick brown fox jumps
XNLI	Sent. Pair Labeling	15	15 → 15	The quick brown fox jumps The fox is fast ↲ Entails
SimAlign	Unsupervised Alignment	–	1 → 6	Le renard brun rapide saute The quick brown fox jumps

Table 1: Summary of the linguistic information we probe XLM-R$_{replica}$ for throughout pretraining.

step (or layer) in pretraining that performs the best across all languages and suggest that methods that better balance these tradeoffs could improve multilingual pretraining in the future.

2 Analyzing Knowledge Acquisition Throughout Multilingual Pretraining

Our goal is to quantify when information is learned by multilingual models across pretraining. To this end, we reproduce a popular multilingual pretrained model, XLM-R – referred to as XLM-R$_{replica}$ – and retain several training checkpoints (§2.1). A suite of linguistic tasks is then run on the various checkpoints (§2.2). For a subset of these tasks, we also evaluate at which layer in the network information is captured during pretraining.

Since we want to identify what knowledge is gleaned from the pretraining signal, each task is evaluated without finetuning. The majority of our tasks are tested via *probes*, in which representations are taken from the final layer of the frozen checkpoint and used as input features to a linear model trained on the task of interest (Belinkov et al., 2020). Additional evaluations we consider for the model include an intrinsic evaluation of model learning (BPC) and unsupervised word alignment of model representations. Each of the tasks in our evaluation suite tests the extent to which a training checkpoint captures some form of *linguistic information*, or a specific aspect of linguistic knowledge, and they serve as a proxy for language understanding in the model.

2.1 Replicating XLM-R

Analyzing model learning throughout pretraining requires access to intermediate training checkpoints, rather than just the final artifact. We replicate the base version of XLM-R and save a number of checkpoints throughout the training process. Our pretraining setup primarily follows that of the original XLM-R, with the exception that we use a smaller batch size (1024 examples per batch instead of 8192) due to computational constraints. All other hyperparameters remain unchanged.

XLM-R$_{replica}$ is also trained on the same data as the original model, CC100. This dataset consists of filtered Common Crawl data for 100 languages, with a wide range of data quantities ranging from 0.1 GiB for languages like Xhosa and Scottish Gaelic to over 300 Gib for English. As with XLM-R, we train on CC100 for 1.5M updates and save 39 checkpoints for our analysis, with more frequent checkpoints taken in the earlier portion of training: we save the model every 5k training steps up to the 50k step, and then every 50k steps. Further details about the data and pretraining scheme can be found in Conneau et al. (2020a). We compare the final checkpoint of XLM-R$_{replica}$ to the original XLM-R$_{base}$ and find that while XLM-R$_{replica}$ performs slightly worse in-language, the two models perform similarly cross-lingually (Appendix A).

2.2 Linguistic Information Tasks

The analysis suite covers different types of syntactic knowledge, semantics in the form of natural language inference, and word alignment (Table 1). These tasks evaluate both in-language linguistics as well as cross-lingual transfer with a wide variety of languages and language pairs. Unless otherwise stated (§5), we evaluate the output from the final layer of XLM-R$_{replica}$. Additionally, most tasks (POS tagging, dependency structure tasks,

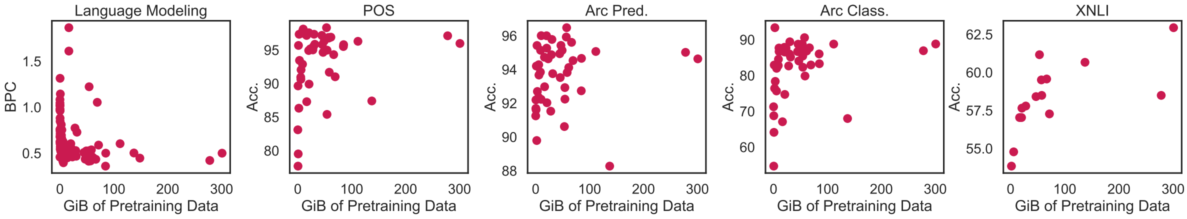

Figure 1: Best in-language performance of XLM-R$_{replica}$ on various tasks and languages across all checkpoints.

and XNLI) are evaluated with accuracy; the MLM evaluation is scored on BPC, and SimAlign is evaluated on F1 performance. Appendix A details the languages covered by each of these tasks and further experimental details.

MLM Bits per Character (BPC) As an intrinsic measure of model performance, we consider the bits per character (BPC) on each training language of the underlying MLM. For a sequence **s**, BPC(**s**) is the (average) negative log-likelihood (NLL) of the sequence under the model normalized by the number of characters per token; lower is better for this metric. These numbers are often not reported for individual languages or across time for multilingual models, making it unclear how well the model captures each language on the pretraining task. We evaluate BPC on the validation split of CC100.

Part-of-Speech (POS) Tagging We probe XLM-R$_{replica}$ with a linear model mapping the representation for each word to its corresponding POS tag; words that are split into multiple subword tokens in the input are represented by the average of their subword representations. The probes are trained using the Universal Dependencies (UD) treebanks for each language (Nivre et al., 2020). For cross-lingual transfer, we evaluate a subset of languages that occur in Parallel Universal Dependencies (PUD; Zeman et al., 2017), a set of parallel test treebanks, to control for any differences in the evaluation data.

Dependency Structure We evaluate syntactic dependency structure knowledge with two pair-wise probing tasks: *arc prediction*, in which the probe is trained to identify pairs of words that are linked with a dependency arc; and *arc classification*, where the probe labels a pair of words with their corresponding dependency relation. The two word-level representations r_1 and r_2 are formatted as a single concatenated input vector $[r_1; r_2; r_1 \odot r_2]$, following Blevins et al. (2018). This combined representation is then used as the input to a linear model that labels the word pair. Probes for both dependency tasks are trained and evaluated with the same set of UD treebanks as POS tagging.

XNLI We also consider model knowledge of natural language inference (NLI), where the probe is trained to determine whether a pair of sentences entail, contradict, or are unrelated to each other. Given two sentences, we obtain their respective representation r_1 and r_2 by averaging all representations in the sentence, and train the probe on the concatenated representation $[r_1; r_2; r_1 \odot r_2]$. We train and evaluate the probes with the XNLI dataset (Conneau et al., 2018); for training data outside of English, we use the translated data provided by Singh et al. (2019).

Word Alignment In the layer-wise evaluation (§5), we evaluate how well the model's internal representations are aligned using SimAlign (Sabet et al., 2020), an unsupervised algorithm for aligning bitext at the word level using multilingual representations. We evaluate the XLM-R$_{replica}$ training checkpoints with SimAlign on manually annotated reference alignments for the following language pairs: EN-CS (Mareček, 2008), EN-DE[2], EN-FA (Tavakoli and Faili, 2014), EN-FR (WPT2003, Och and Ney, 2000), EN-HI[3], and EN-RO[3].

3 In-language Learning Throughout Pretraining

We first consider the in-language, or monolingual, performance of XLM-R$_{replica}$ on different types of linguistic information across pretraining. We find that in-language linguistics is learned (very) early in pretraining and is acquired in a consistent order, with lower-level syntactic information learned before more complex syntax and semantics. Additionally, the final checkpoint of XLM-R$_{replica}$ often

[2] Gold alignments on EuroParl (Koehn, 2005), http://www-i6.informatik.rwth-aachen.de/goldAlignment/

[3] WPT2005, http://web.eecs.umich.edu/ mihalcea/wpt05/

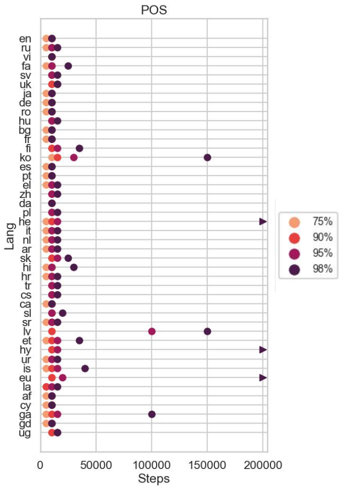

Figure 2: Learning progress of XLM-R$_{replica}$ on POS tagging, up to 200k training steps. Each point represents the step at which the model achieves x% of the best overall performance of the model on that task; arrows indicate languages that reach the 98% mark after 200k steps.

experiences performance degradation compared to the best checkpoint for a language, suggesting that the model is forgetting information for a number of languages by the end of pretraining.

3.1 Monolingual Performance for Different Languages

Figure 1 presents the overall best performance of the model across time on the considered tasks and languages. We observe a large amount of variance in performance on each task. Across languages, XLM-R$_{replica}$ performance ranges between 1.86 and 0.36 BPC for language modeling, 88.3% and 96.5% accuracy for dependency arc prediction, 77.67% and 98.3% accuracy for POS tagging, 54.7% and 93.3% accuracy for arc classification, and 53.8% and 62.9% accuracy for XNLI. Overall, these results confirm previous findings that multilingual model performance varies greatly on different languages (§6).

3.2 When Does XLM-R Learn Linguistic Information?

Figure 2 shows the step at which XLM-R$_{replica}$ reaches different percentages of its best performance of the model on POS tagging. Figures for the other tasks are given in Appendix D.

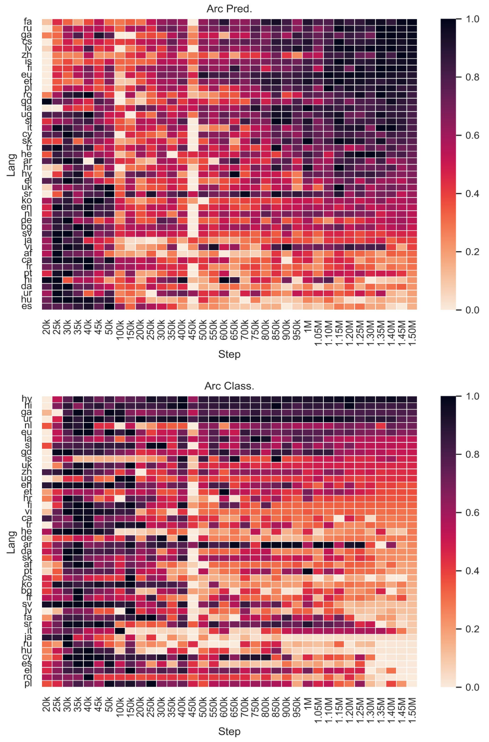

Figure 3: Heatmap of relative performance over time for dependency arc prediction and classification. Languages are ordered by performance degradation in the final training checkpoint.

Monolingual linguistics is acquired early in pretraining We find that XLM-R$_{replica}$ acquires the majority of in-language linguistic information early in training. However, the average time step for acquisition varies across tasks. For dependency arc prediction, all languages achieve 98% or more of total performance by 20k training steps (out of 1.5M total updates). In contrast, XNLI is learned later with the majority of the languages achieving 98% of the overall performance after 100k training updates. This order of acquisition is in line with monolingual English models, which have also been found to learn syntactic information before higher-level semantics (Liu et al., 2021).

We also observe that this order of acquisition is often maintained within individual languages. 12 out of 13 of the languages shared across all tasks reach 98% of the best performance consistently in the order of POS tagging and arc prediction (which are typically learned within one checkpoint of each other), arc classification, and XNLI.

Model behavior later in pretraining varies across languages For some languages and tasks, XLM-R$_{replica}$ never achieves good absolute performance (Figure 1). For others, the performance of XLM-R$_{replica}$ decreases later in pretraining, leading the converged model to have degraded performance on those tasks and languages (Figure 3).

We hypothesize that this is another aspect of the “curse of multilinguality,” where some languages are more poorly captured in multilingual models due to limited model capacity (Conneau et al., 2020a; Wang et al., 2020), that arises during the training process. We also find that the ranking of languages by performance degradation is not correlated across tasks. This suggests the phenomenon is not limited to a subset of low-resource languages and can affect any language learned by the model.

More generally, these trends demonstrate that the best model state varies across languages and tasks. Since BPC continues to improve on all individual training languages throughout pretraining (Appendix D), the results also indicate that performance on the pretraining task is not directly tied to performance on the linguistic probes. This is somewhat surprising, given the general assumption that better pretraining task performance corresponds to better downstream task performance.

4 Cross-lingual Transfer Throughout Pretraining

Another question of interest is: when do multilingual models learn to transfer between languages? We find that cross-lingual transfer is acquired later in pretraining than monolingual linguistics and that the step at which XLM-R$_{replica}$ learns to transfer a specific language pair varies greatly. Furthermore, though the order in which XLM-R$_{replica}$ learns to transfer different linguistic information across languages is on average consistent with in-language results, the order in which the model learns to transfer across specific language pairs for different tasks is much more inconsistent.

4.1 Overall Transfer Across Language Pairs

Which languages transfer well? Figure 4 shows cross-lingual transfer between different language pairs; most source languages perform well in-language (the diagonal). We observe that some tasks, specifically dependency arc prediction, are easier to transfer between languages than others; however, across the three tasks with shared language pairs (POS tagging, arc prediction, and arc classification) we see similar behavior in the extent to which each language transfers to others. For example, English and Italian both transfer well to most of the target languages. However, other languages are isolated and do not transfer well into or out of other languages, even though in some cases, such as Japanese, the model achieves good in-language performance.

On XNLI there is more variation in in-language performance than is observed on the syntactic tasks. This stems from a more general trend that some languages appear to be easier to transfer into than others, leading to the observed performance consistency within columns. For example, English appears to be particularly easy for XLM-R$_{replica}$ to transfer into, with 12 out of the 14 non-English source languages performing as well or better on English as in-language.

Cross-lingual transfer is asymmetric We also find that language transfer is asymmetric within language pairs (Figure 5). There are different transfer patterns between dependency arc prediction and the other syntactic tasks: for example, we see that Korean is worse relatively as a source language than as the target for POS tagging and arc classification, but performs better when transferring to other languages in arc prediction. However, other languages such as Arabic have similar trends across the syntactic tasks. On XNLI, we find that Swahili and Arabic are the most difficult languages to transfer into, though they transfer to other languages reasonably well.

These results expand on observations in Turc et al. (2021) and emphasize that the choice of source language has a large effect on cross-lingual performance in the target. However, there are factors in play in addition to linguistic similarity causing this behavior, leading to asymmetric transfer within a language pair. We further examine these correlations with overall cross-lingual performance and asymmetric transfer in §B.2.

4.2 When is Cross-lingual Transfer Learned During Pretraining?

We next consider when during pretraining XLM-R$_{replica}$ learns to transfer between languages (Figure 6; the dotted line indicates the 200k step cutoff used in Figure 2 for comparison). Unlike the case of monolingual performance, the step at which the model acquires most cross-lingual signal (98%) varies greatly across language pairs. We also find that (similar to the in-language setting) higher-level linguistics transfer later in pretraining than lower-level ones: the average step for a language pair to achieve 98% of overall performance occurs at 115k for dependency arc prediction, 200k for POS tagging, 209k for dependency arc classification,

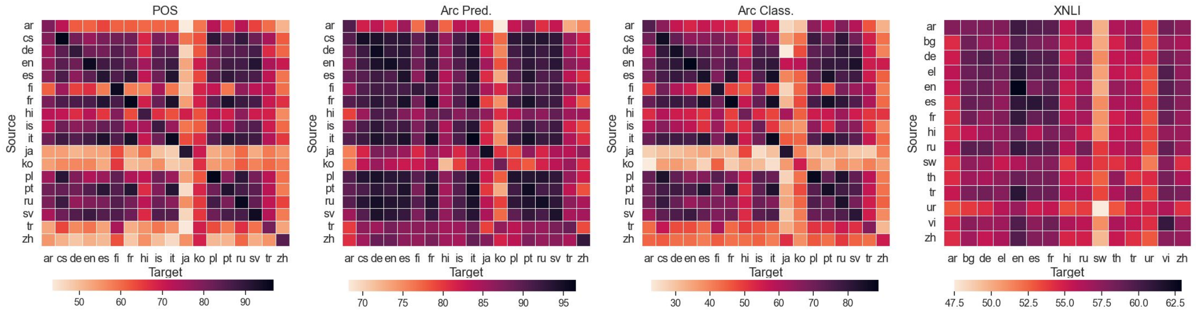

Figure 4: Overall performance of XLM-R$_{replica}$ on each analysis task when transferring from various source to target languages.

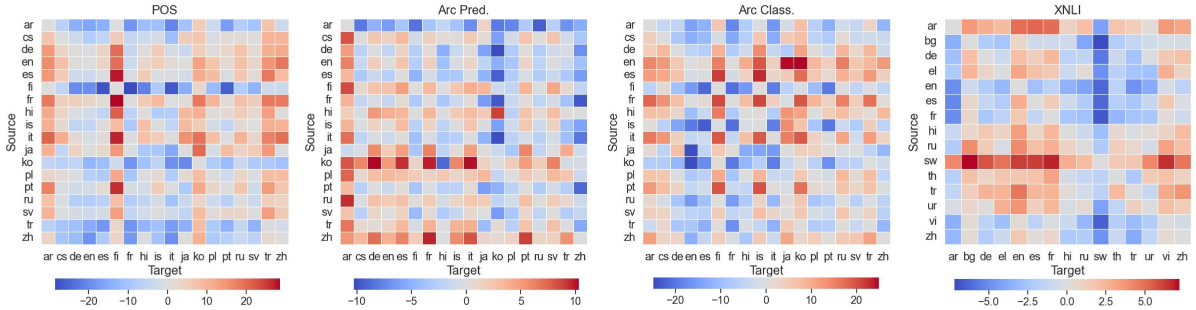

Figure 5: Heatmap of the asymmetry of cross-lingual transfer in XLM-R$_{replica}$. Each cell shows the difference in performance between language pairs ($l_1 \to l_2$) and ($l_2 \to l_1$).

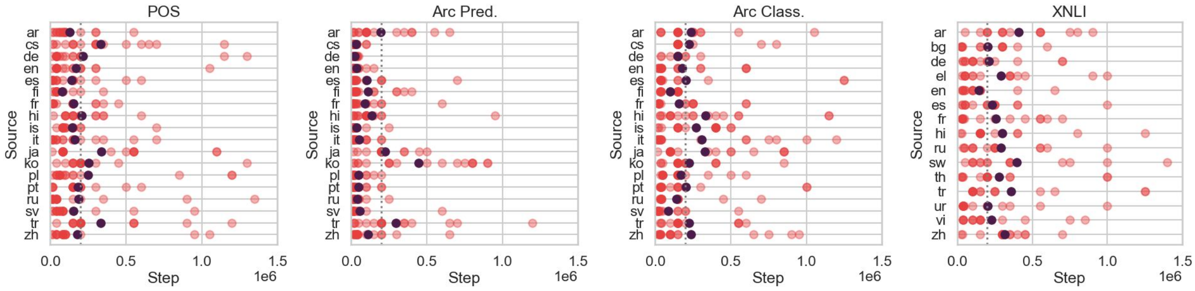

Figure 6: Cross-lingual learning progress of XLM-R$_{replica}$ across pretraining. Each red point represents the step to 98% of the best performance for a language pair; the purple represents the mean 98% transfer step for the source language.

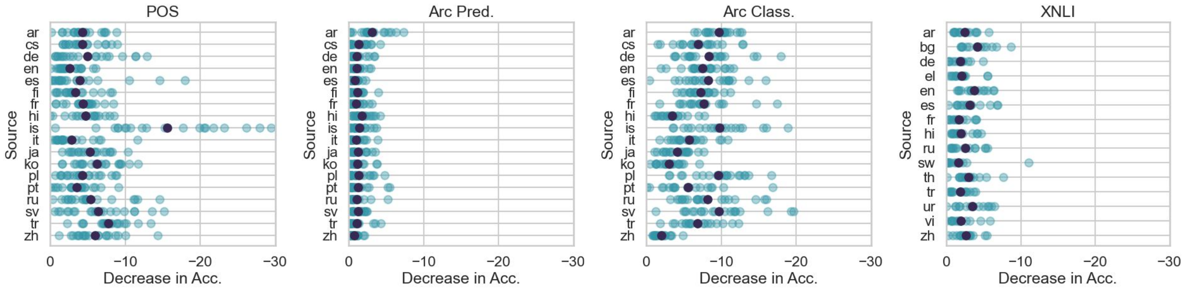

Figure 7: Degradation of cross-lingual transfer performance of XLM-R$_{replica}$ across pretraining. Each blue point represents the change in performance from the overall best step to the final model checkpoint for a language pair; the navy represents the mean decrease for the source language.

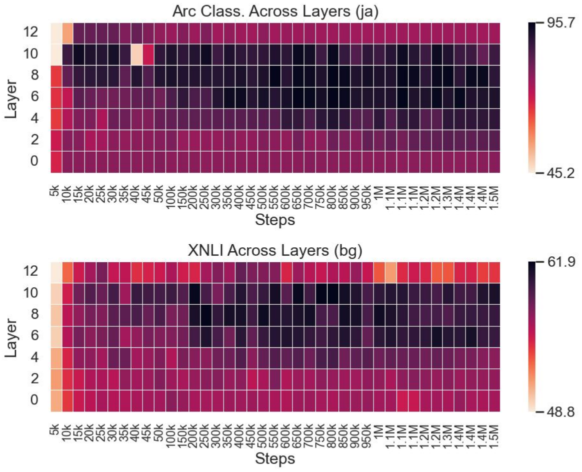

Figure 8: Heatmap of XLM-R$_{replica}$ performance for Japanese arc classification and Bulgarian XNLI. Additional heatmaps are given in Appendix C.

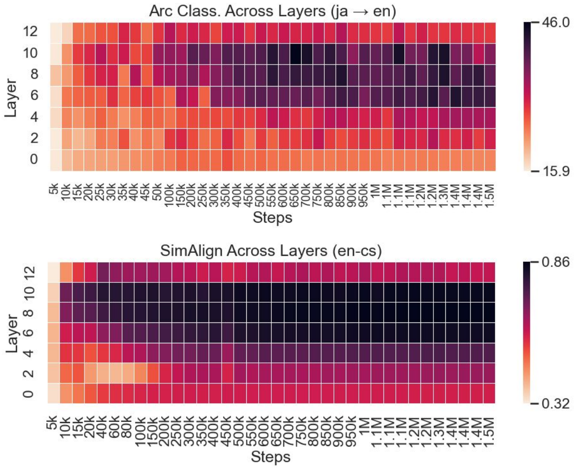

Figure 9: Heatmap of XLM-R$_{replica}$ cross-lingual performance by layer for arc classification (JA → EN) and SimAlign (EN-CS).

and 274k for XNLI. In contrast, when the model learns to transfer different linguistic information between two specific languages can vary wildly: only approximately 21% of the language pairs shared across the four tasks transfer in the expected order.

We also investigate the amount to which the cross-lingual abilities of XLM-R$_{replica}$ decrease over time (Figure 7; more detailed across time results for transferring out of English are given in Appendix D). Similarly to in-language behavior, we find that the model exhibits notable performance degradation for some language pairs (in particular on POS tagging and dependency arc classification), and the extent of forgetting can vary wildly across target languages for a given source language.

5 Layer-wise Learning Throughout Pretraining

In the experiments above we show that in many cases the final layer of XLM-R$_{replica}$ forgets information by the end of pretraining. Motivated by this, we investigate whether this information is retained in a different part of the network by probing how information changes *across* layers during pretraining. We find a surprising trend in how the best-performing layer changes over time: the model acquires knowledge in higher layers early on, which then propagates to and improves in the lower layers later in pretraining.

5.1 In-language Knowledge Across Layers

We first look at the layer-wise performance of XLM-R$_{replica}$ on a subset of languages for dependency arc classification (CS, EN, HI, and JA) and XNLI (BG, EN, HI, and ZH) (Figure 8). We find that the last layer is often not the best one for each task, with lower layers often outperforming the final one. On average, the best internal layer state outperforms the final layer of XLM-R$_{replica}$ by 7.59 accuracy points on arc classification and 2.93 points on XNLI.

We also observe a trend of lower layers acquiring knowledge later in training than the final one. To investigate this, we calculate the expected best layer (i.e., the average layer weighted by performance) at each checkpoint and find that it decreases over time, by up to 2.79 layers for arc classification and 2.49 layers for XNLI (Appendix C), indicating that though the final layer quickly fits to the forms of in-language information we test for, this information then shifts to lower layers in the network over time.

5.2 Cross-lingual Knowledge Across Layers

Next, we consider how cross-lingual transfer skills are captured across layers during pretraining. Every other XLM-R$_{replica}$ layer is evaluated on the subsets of languages for arc classification and XNLI in §5.1. We also use SimALign to test how well word representations at these layers align from English to {CS, DE, FA, FR, HI, RO}.We observe similar trends with respect to layer performance over time to the in-language results (Figure 9; additional results given in Appendix C). Specifically, we observe an average decrease in the expected layer of 1.10 (ranging from 0.67 to 2.20) on arc classification, 1.02 (ranging from 0.37 to 2.01) on XNLI, and 1.66 (ranging from 0.83 to 2.41) on SimAlign.

We also observe that while most layers perform relatively well in-language performance, the lowest layers of XLM-R$_{replica}$ (layers 0-4) often perform much worse than the middle and final layers for cross-lingual transfer throughout the pretraining process – for example, in the case of Japanese to English on arc classification. We hypothesize that this is due to better alignment across languages in later layers, similar to the findings in Muller et al. (2021).

6 Related Work

Linguistic knowledge in multilingual models There have been several different approaches to quantifying the linguistic information that is learned by multilingual models. One direction has performed layer-wise analyses to quantify what information is stored at different layers in the model (de Vries et al., 2020; Taktasheva et al., 2021; Papadimitriou et al., 2021). Others have examined the extent to which the different training languages are captured by the model, finding that some languages suffer in the multilingual setting despite the overall good performance exhibited by the models (Conneau et al., 2020a; Wang et al., 2020).

Cross-lingual transfer in multilingual models Another line of analysis seeks to understand the cross-lingual abilities of multilingual models. Chi et al. (2020) show that subspaces of mBERT representations that capture syntax are approximately shared across languages, suggesting that portions of the model are cross-lingually aligned. A similar direction of interest is whether multilingual models learn language-agnostic representations. Singh et al. (2019) find that mBERT representations can be partitioned by language, indicating that the representations retain language-specific information. Similarly, other work has shown that mBERT representations can be split into language-specific and language-neutral components (Libovickỳ et al., 2019; Gonen et al., 2020; Muller et al., 2021).

Other work has investigated the factors that affect cross-lingual transfer. These factors include the effect of sharing subword tokens on cross-lingual transfer (Conneau et al., 2020b; K et al., 2020; Deshpande et al., 2021) and which languages act as good source languages for cross-lingual transfer (Turc et al., 2021). Notably, Lauscher et al. (2020), K et al. (2020) and Hu et al. (2020) find that multilingual pretrained models perform worse when transferring to distant languages and low-resource languages.

Examining Pretrained Models Across Time A recent direction of research has focused on probing multiple checkpoints taken from different points in the pretraining process, in order to quantify when the model learns information. These works have examined the acquisition of syntax (Pérez-Mayos et al., 2021) as well as higher-level semantics and world knowledge over time (Liu et al., 2021) from the RoBERTa pretraining process. Similarly, Chiang et al. (2020) perform a similar temporal analysis for AlBERT, and Choshen et al. (2022) find that the order of linguistic acquisition during language model training is consistent across model sizes, random seeds, and LM objectives.

Most work on probing pretrained models across the training process has focused on monolingual, English models. There are some limited exceptions: Dufter and Schütze (2020) present results for multilingual learning in a synthetic bilingual setting, and Wu and Dredze (2020) examine performance across pretraining epochs for a small number of languages. However, this paper is the first to report a comprehensive analysis of monolingual and cross-lingual knowledge acquisition on a large-scale multilingual model.

7 Conclusion

In this paper, we probe training checkpoints across time to analyze the training dynamics of the XLM-R pretraining process. We find that although the model learns in-language linguistic information early in training – similar to findings on monolingual models – cross-lingual transfer is obtained all throughout the pretraining process.

Furthermore, the order in which linguistic information is acquired by the model is generally consistent, with lower-level syntax acquired before semantics. However, we observe that for individual language pairs this order can vary wildly, and our statistical analyses demonstrate that model learning speed and overall performance on specific languages (and pairs) are difficult to predict from language-specific factors.

We also observe that the final model artifact of XLM-R$_{replica}$ performs often significantly worse than earlier training checkpoints on many languages and tasks. However, layer-wise analysis of the model shows that linguistic information shifts lower in the network during pretraining, with lower layers eventually outperforming the final layer. Altogether, these findings provide a better understanding of multilingual training dynamics that can inform future pretraining approaches.

8 Limitations

We note some potential limitations of this work. We consider a single pretraining setting (replicating the training of XLM-R$_{base}$), and the extent to which our findings transfer to other multilingual pretraining settings remains an open question. In particular, pretraining a language model with more parameters or on different multilingual data could lead to other trends, though many of our findings are consistent with prior work.

Additionally, despite our attempts to use diverse datasets for evaluating these models, the language choices available in annotated NLP data are skewed heavily towards Indo-European, especially English and other Western European, languages. This means that many of the low-resource languages seen in the pretraining data are unaccounted for in this study. Due to this, we only evaluate word alignment between six languages paired with English, and a number of the non-English datasets we use are translated from English.

Another product of limited multilingual resources is our ability to compare across languages; in UD, each treebank is annotated on different domains with different dataset sizes. This limits the comparisons we can make across probe training settings, though we focus on changes within individual languages in this work. To address this limitation, we use the parallel test sets from Parallel Universal Dependencies for our cross-lingual transfer experiments, which allows us to compare performance on different target languages from the same source language directly.

Acknowledgements

We would like to thank Naman Goyal for his advice on retraining XLM-R and Masoud Jalili Sabet for providing the data to run SimAlign. We also thank Leo Z. Liu for helpful conversations on this work as well as the anonymous reviewers for their thoughtful feedback.

References

Yonatan Belinkov, Sebastian Gehrmann, and Ellie Pavlick. 2020. Interpretability and analysis in neural NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 1–5, Online. Association for Computational Linguistics.

Terra Blevins, Omer Levy, and Luke Zettlemoyer. 2018. Deep RNNs encode soft hierarchical syntax. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–19.

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

Cheng-Han Chiang, Sung-Feng Huang, and Hung-Yi Lee. 2020. Pretrained language model embryology: The birth of albert. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6813–6828.

Leshem Choshen, Guy Hacohen, Daphna Weinshall, and Omri Abend. 2022. The grammar-learning trajectories of neural language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8281–8297, Dublin, Ireland. Association for Computational Linguistics.

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

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485.

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.

Wietse de Vries, Andreas van Cranenburgh, and Malvina Nissim. 2020. What's so special about bert's layers? a closer look at the nlp pipeline in monolingual and multilingual models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4339–4350.

Jacob Delvin. 2019. Multilingual BERT Readme. `https://github.com/google-research/bert/blob/master/multilingual.md`.

Ameet Deshpande, Partha Talukdar, and Karthik Narasimhan. 2021. When is bert multilingual? isolating crucial ingredients for cross-lingual transfer. *arXiv preprint arXiv:2110.14782*.

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 (EMNLP)*, pages 4423–4437.

Hila Gonen, Shauli Ravfogel, Yanai Elazar, and Yoav Goldberg. 2020. It's not greek to mbert: Inducing word-level translations from multilingual bert. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 45–56.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

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*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *The Tenth Machine Translation Summit Proceedings of Conference*, pages 79–86. International Association for Machine Translation.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

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

Jindřich Libovickỳ, Rudolf Rosa, and Alexander Fraser. 2019. How language-neutral is multilingual bert? *arXiv preprint arXiv:1911.03310*.

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

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 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Zeyu Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A Smith. 2021. Probing across time: What does roberta know and when? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 820–842.

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Copenhagen, Denmark.

David Mareček. 2008. *Automatic Alignment of Tectogrammatical Trees from Czech-English Parallel Corpus*. Charles University, MFF UK.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021. First align, then predict: Understanding the cross-lingual ability of multilingual bert. In *EACL 2021-The 16th Conference of the European Chapter of the Association for Computational Linguistics*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajic, 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.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 440–447.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Isabel Papadimitriou, Ethan A Chi, Richard Futrell, and Kyle Mahowald. 2021. Deep subjecthood: Higher-order grammatical features in multilingual bert. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2522–2532.

Tommaso Pasini, Alessandro Raganato, Roberto Navigli, et al. 2021. Xl-wsd: An extra-large and cross-lingual evaluation framework for word sense disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press.

Laura Pérez-Mayos, Miguel Ballesteros, and Leo Wanner. 2021. How much pretraining data do language models need to learn syntax? In *Proceedings of the*

2021 Conference on Empirical Methods in Natural Language Processing, pages 1571–1582.

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.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. Simalign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643.

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

Ekaterina Taktasheva, Vladislav Mikhailov, and Ekaterina Artemova. 2021. Shaking syntactic trees on the sesame street: Multilingual probing with controllable perturbations. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 191–210.

Leila Tavakoli and Heshaam Faili. 2014. Phrase alignments in parallel corpus using bootstrapping approach. *International Journal of Information and Communication Technology Research.*

Iulia Turc, Kenton Lee, Jacob Eisenstein, Ming-Wei Chang, and Kristina Toutanova. 2021. Revisiting the primacy of english in zero-shot cross-lingual transfer. *CoRR*.

Zirui Wang, Zachary C Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual bert? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, et al. 2017. Conll 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19. Association for Computational Linguistics.

A Linguistic Probe Details

Table 2 presents the languages that are included in each of the probing tasks. We filter the Romanized versions of languages from the CC100 dataset, leaving us with 94 for evaluation.

A.1 Experimental Setup

Each evaluation is run on the frozen parameters of a training checkpoint of XLM-R$_{replica}$. All representations are taken from the final (12th) layer of the encoder, except for the experiments presented in §5, which consider the performance of different layers within the model over time.

For the linguistic information tasks involving probing, each probe consists of a single linear layer, trained with a batch size of 256 for 50 epochs with early stopping performed on the validation set. The probes therefore consist of a limited number of parameters $m * l$, where $m = 768$ is the output dimension of the model and l is the size of the task label set. Following Liu et al. (2019), the probes are optimized with a learning rate of 1e-3. Each probe is trained on a single Nvidia V100 16GB GPU and takes between <1 minute and 6 minutes to train (depending on dataset size, which varies by language and task). The reported results for each probe are the averaged performance across five runs.

For SimAlign, we use the default settings provided in the SimAlign implementation.[4] We report word-level alignment performance (instead of sub-word alignment) using the itermax alignment algorithm.

A.2 XLM-R Replication Details

XLM-R$_{replica}$ consists of the same model architecture as XLM-R$_{base}$, with a total of 270M parameters. We train the model for 1.5 million updates on 64 Nvidia V100 32 GB GPUs using the fairseq

Task	Languages
BPC	af, am, ar, as, az, be, bg, bn, br, bs, ca, cs, cy, da, de, el, en, eo, es, et, eu, fa, fi, fr, fy, ga, gd, gl, gu, ha, he, hi, hr, hu, hy, id, is, it, ja, jv, ka, kk, km, kn, ko, ku, ky, la, lo, lt, lv, mg, mk, ml, mn, mr, ms, my, ne, nl, no, om, or, pa, pl, ps, pt, ro, ru, sa, sd, si, sk, sl, so, sq, sr, su, sv, sw, ta, te, th, tl, tr, ug, uk, ur, uz, vi, xh, yi, zh, zh
UD	af, **ar**, bg, ca, **cs**, cy, da, **de**, el, **en**, **es**, et, eu, fa, **fi**, **fr**, ga, gd, he, **hi**, hr, hu, hy, **is**, **it**, **ja**, **ko**, la, lv, nl, **pl**, **pt**, ro, **ru**, sk, sl, sr, **sv**, **tr**, ug, uk, ur, vi, **zh**
XNLI	**ar**, **bg**, **de**, **el**, **en**, **es**, **fr**, **hi**, **ru**, **sw**, **th**, **tr**, **ur**, **vi**, **zh**

Table 2: Table summarizing the languages considered for each task. Languages in bold are also used for the cross-lingual setting of the task. UD covers all of the languages used for POS tagging, dependency arc prediction, and dependency arc classification.

[4] https://github.com/cisnlp/simalign

Task		XLM-R$_{base}$	XLM-R$_{replica}$
In-lang	BPC	0.609*	0.652
	POS	89.65*	87.20
	XNLI	58.08*	55.73
X-lang	POS	66.01*	64.94
	XNLI	53.26	53.77*

Table 3: Average performance across languages of XLM-R$_{base}$ and the final checkpoint of XLM-R$_{replica}$.

library (Ott et al., 2019). Notably, the language sampling alpha for up-weighting less frequent languages is set to $\alpha = 0.7$: this matches the value used for the XLM-R, though it was reported as $\alpha = 0.3$ in the original paper.

A.3 Comparison with XLM-R$_{base}$

We also compare the performance of our retrained XLM-R$_{replica}$ model against the original XLM-R$_{base}$ on a subset of the tasks in our evaluation suite (Table 3). We find that on average, the original XLM-R model achieves better BPC than the replicated model; this is likely due to the decrease in batch size while retraining the model. The replica model also performs slightly worse than the original on in-language tasks but comparably cross-lingually (and outperforms the original model on cross-lingual XNLI).

B What Factors Affect Multilingual Learning?

This section presents extended results analyzing the correlations between different factors and the in-language and cross-lingual learning exhibited by XLM-R$_{replica}$.

B.1 In-language Correlation Study

We consider whether the following factors correlate with various measures of model learning (Table 4): *pretraining data*, the amount of text in the CC100 corpus for each language; *task data*, the amount of in-task data used to train each probe; and *language similarity* to English, which is the highest-resource language in the pretraining data. We use the syntactic distances calculated in Malaviya et al. (2017) as our measure of language similarity; these scores are smaller for more similar language pairs.

Overall Performance The amount of pretraining data and in-task training data are strongly correlated with overall task performance for most of the considered tasks; this corroborates similar results from Wu and Dredze (2020). Language similarity with English is also correlated with better in-task performance on all tasks except for dependency arc prediction, suggesting that some form of cross-lingual signal supports in-language performance for linguistically similar languages.

Learning Progress Measures We also consider (1) the step at which XLM-R$_{replica}$ achieves 95% of its best performance for each language and task, which measures how quickly the model obtains a majority of the tested linguistic information, and (2) how much the model *forgets* from the best performance for each language by the final training checkpoint. We find that language similarity to English is strongly correlated with how quickly XLM-R$_{replica}$ converges on BPC and dependency arc classification. This suggests that cross-lingual signal helps the model more quickly learn lower-resource languages on these tasks, in addition to improving overall model performance. However, we observe no strong trends as to what factors affect forgetting across tasks.

B.2 Cross-lingual Correlation Study

Table 5 presents a correlation study of different measures for cross-lingual transfer in XLM-R$_{replica}$. We consider the effect of source and target pretraining data quantity, the amount of in-task training data (in the source language), and the similarity between the source and target language on the following transfer measures: overall task performance, asymmetry in transfer (the difference in model performance on $l_1 \rightarrow l_2$ compared to $l_2 \rightarrow l_1$), the step at which the model achieves 95% or more of overall performance on the language pair, and forgetting – the (relative) degradation of overall performance in the final model checkpoint.

Correlations of Transfer with Language Factors For overall cross-lingual performance, we observe that language similarity is highly correlated with task performance for all tasks and is similarly correlated with speed of acquisition (the step to 95% of overall performance) for three of the four considered tasks. This is in line with prior work that has also identified language similarity as a strong indicator of cross-lingual performance (Pires et al., 2019). However, all considered factors are less correlated with the other measures of knowledge acquisition, such as the asymmetry of transfer and the forgetting of cross-lingual knowledge; this sug-

Variable	Factors	Spearman (ρ)				
		BPC	POS	Arc Pred.	Arc Class.	XNLI
Task Perf.	Pretraining Data	-0.597**	0.258	0.267	0.411*	0.767**
	Task Data	-0.597**	0.462*	0.276	0.527**	
	Lang Sim.	0.427**	-0.315*	-0.170	-0.427*	-0.779**
Steps to 95%	Pretraining Data	0.135	-0.290	-0.193	-0.301*	-0.239
	Task Data	0.135	-0.065	-0.260	-0.209	
	Lang Sim.	-0.385**	0.156	0.268	0.325*	0.316
Forgetting	Pretraining Data		0.230	0.218	0.437*	0.564*
	Task Data		-0.322*	-0.338*	-0.015	
	Lang Sim.		0.172	-0.158	-0.181	-0.795**

Table 4: Correlation study of different factors against measures of in-language knowledge. * $p < 0.05$, ** $p < 0.001$

Variable	Factors	Spearman (ρ)			
		POS	Arc Pred.	Arc Class.	XNLI
Task Perf.	Src. Pretraining Data	0.113*	0.107	0.117*	0.178*
	Trg. Pretraining Data	0.038	0.144*	0.015	0.625**
	Task Data	0.245**	0.124*	0.129*	
	Lang Sim.	-0.598**	-0.575**	-0.593**	-0.321**
Asymmetry	Src. Pretraining Data	0.116*	-0.045	0.140*	-0.423*
	Trg. Pretraining Data	-0.116*	0.045	-0.140*	0.423*
	Task Data	0.123*	-0.016	-0.077	
Steps to 95%	Src. Pretraining Data	-0.290**	-0.023	-0.132*	-0.195*
	Trg. Pretraining Data	-0.123*	-0.066	-0.106	-0.057
	Task Data	0.073	-0.057	0.115*	
	Lang Sim.	0.475**	0.518**	0.492**	0.076
Forgetting	Src. Pretraining Data	-0.208**	-0.123*	0.000	0.137*
	Trg. Pretraining Data	0.042	0.015	0.122*	-0.079
	Task Data	0.009	-0.004	0.078	
	Lang Sim.	0.165*	0.186*	-0.025	0.164*

Table 5: Correlation study of different factors against measures of cross-lingual transfer. * $p < 0.05$, ** $p < 0.001$

gests that there could be other factors that explain these phenomena.

Interactions Between Learning Measures We also consider the correlations between the different measures of model performance on cross-lingual transfer. For example, overall transfer performance is strongly correlated (p«0.001) with earlier acquisition (step to 95% of overall performance) for all syntactic tasks: $\rho = -0.50$ for both POS tagging and dependency arc prediction and -0.55 for arc classification. To a lesser extent, overall transfer performance and model forgetting are negatively correlated, ranging from $\rho = -0.13$ to -0.42 across considered tasks. This indicates that XLM-R$_{replica}$ forgets less of the learned cross-lingual signal for better-performing language pairs, at the expense of already less successful ones.

C Expanded Layer-wise Analysis

This section expands on the layer-wise analysis of XLM-R$_{replica}$ presented in §5. Figure 10 gives additional layer-wise heatmaps over time. Figure 11 shows the expected layer (i.e., average layer weighted by relative performance) of XLM-

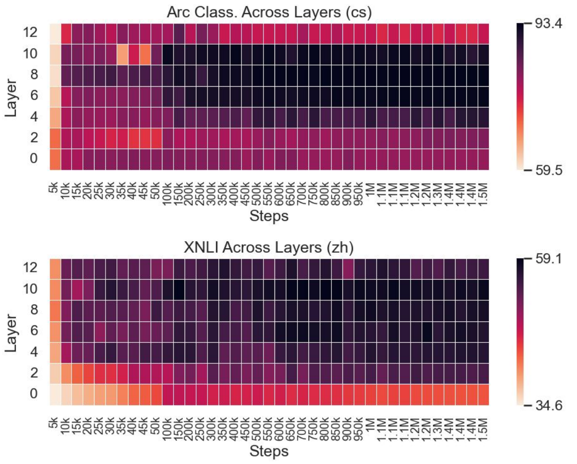

Figure 10: Layer-wise performance heatmaps for Czech arc classification and Chinese XNLI.

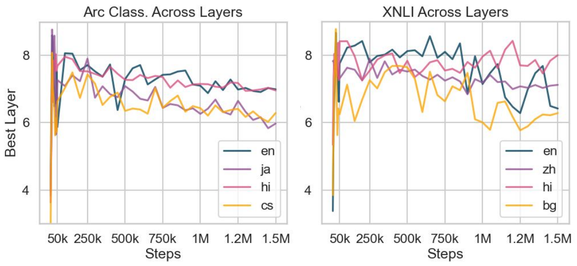

Figure 11: The expected best layer for in-language dependency arc classification and XNLI over time on XLM-R$_{replica}$.

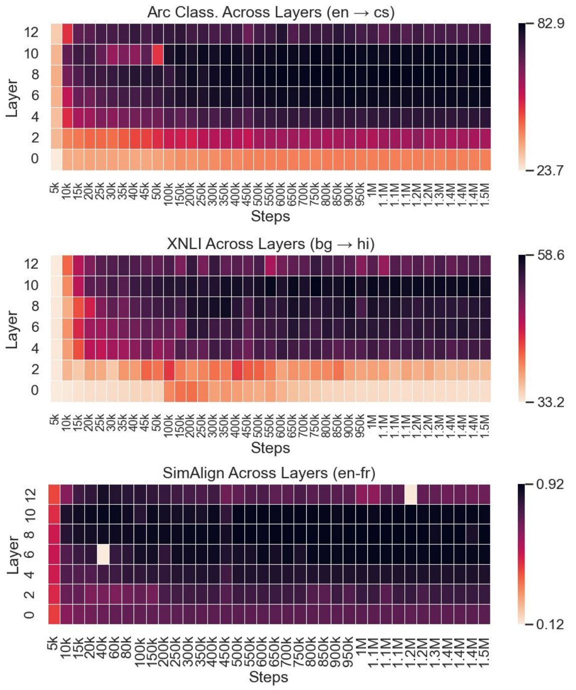

Figure 12: Additional heatmaps of cross-lingual transfer at different layers and timesteps of XLM-R$_{replica}$.

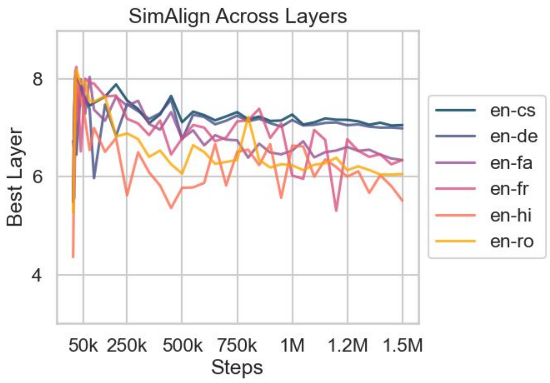

Figure 13: Change in the expected best layer for word alignment via SimAlign over time in XLM-R$_{replica}$

R$_{replica}$ at different time steps. The expected layer decreases over time: by 1.79, 1.61, 1.08, and 2.79 for CS, EN, HI, and JA respectively on dependency arc classification; and by 2.49, 2.25, 0.43, and 0.77 for BG, EN, HI, and ZH respectively on XNLI.

We also provide additional examples of layer-wise cross-lingual transfer in Figure 12; we find that for cross-lingual transfer, the best internal layer outperforms the best final layer state on average by 7.67 on arc classification transfer, 3.39 on XNLI, and 14.2 F1 on Simalign. Figure 13 shows the change in the expected best layer over time for SimAlign.

D Additional Across Time Analyses

This section includes additional results from our analysis of knowledge acquisition during multilingual pretraining:

- Figure 14 presents BPC learning curves for each language in the CC100 training data.
- Figure 16 covers the learning progress of XLM-R$_{replica}$ on dependency arc prediction, arc classification, and XNLI, expanding on the results in §3.2.
- Figure 15 gives the relative performance for in-language POS and XNLI across training checkpoints discussed in §3.2.
- Figure 17 presents more detailed results for relative performance over time when transferring out of English. This expands on the summary figures discussed in §4.2.

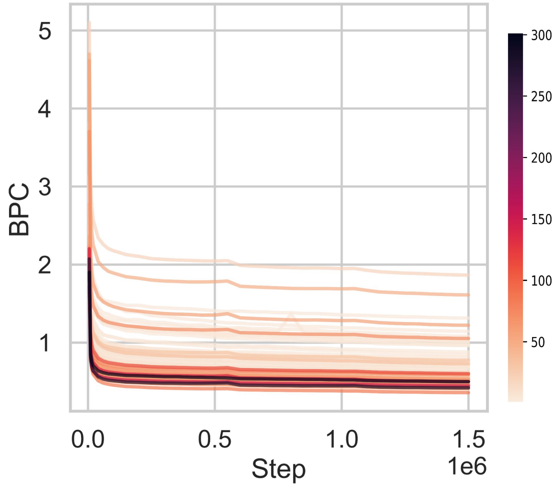

Figure 14: Learning Curves for BPC in each training language. Lines are colored by the amount of pretraining data available for that language.

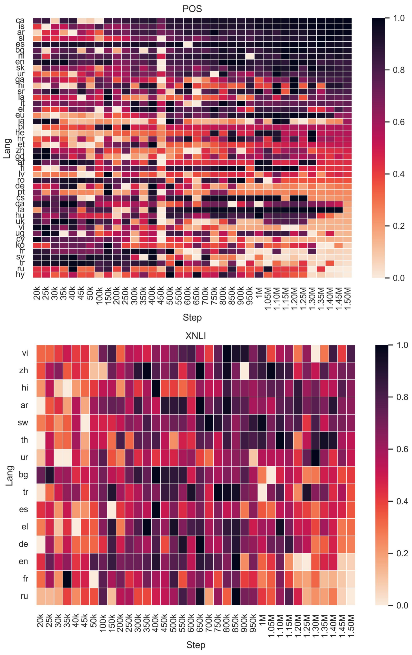

Figure 15: Heatmap of relative performance over time for different languages for POS tagging and XNLI. Languages are ordered by the amount of performance degradation at the final checkpoint.

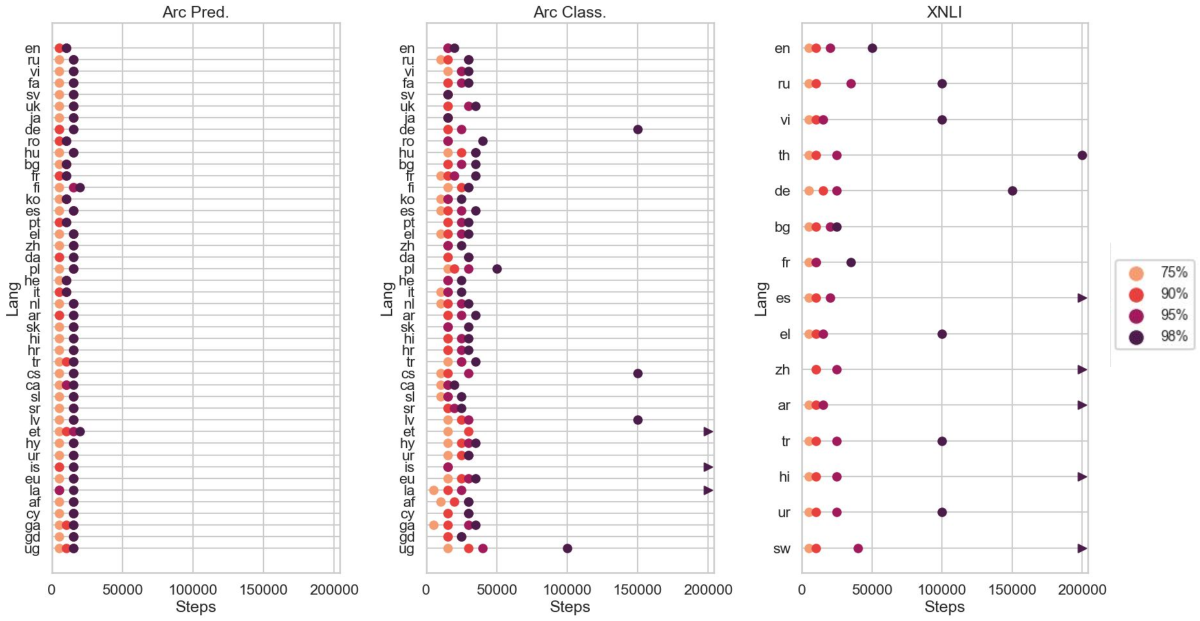

Figure 16: Learning Progress of XLM-R$_{replica}$ across training, up to 200k training steps. Each point represents the step at which the model achieves x% of the best overall performance of the model on that task.

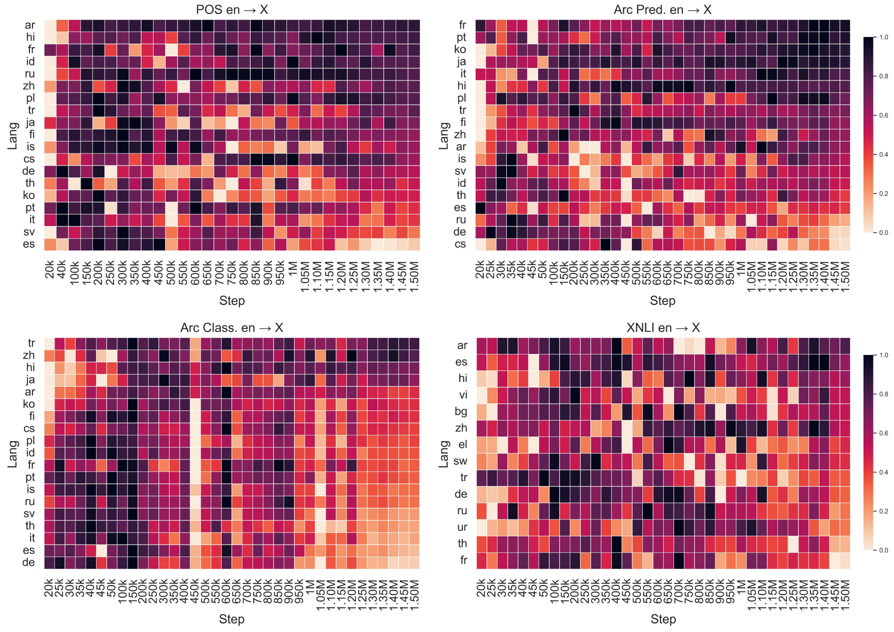

Figure 17: Heatmap of relative performance over time for cross-lingual transfer with English as the source language. Languages are ordered by the amount of performance degradation at the final checkpoint.