Where Does Linguistic Information Emerge in Neural Language Models? Measuring Gains and Contributions Across Layers

Jenny Kunz and **Marco Kuhlmann**
Dept. of Computer and Information Science
Linköping University
jenny.kunz@liu.se and marco.kuhlmann@liu.se

Abstract

Probing studies have extensively explored where in neural language models linguistic information is located. The standard approach to interpreting the results of a probing classifier is to focus on the layers whose representations give the highest performance on the probing task. We propose an alternative method that asks where the task-relevant information emerges in the model. Our framework consists of a family of metrics that explicitly model local information gain relative to the previous layer and each layer's contribution to the model's overall performance. We apply the new metrics to two pairs of syntactic probing tasks with different degrees of complexity and find that the metrics confirm the expected ordering only for one of the pairs. Our local metrics show a massive dominance of the first layers, indicating that the features that contribute the most to our probing tasks are not as high-level as global metrics suggest.

1 Introduction

Probing neural language models aims at finding evidence of learned linguistic structure in the models' parameters by empirically testing hypotheses about the learned representations (Hupkes et al., 2018; Alain and Bengio, 2017). This is often done by training a probing classifier on a diagnostic task with the representations at different layers as the input, and comparing task performance across layers. While probes are conceptually simple and widely used, the methodology and in particular the interpretation of the obtained results is subject to ongoing discussion (Belinkov, 2022).

A classical pattern we often see when plotting probing accuracy across layers is that for higher-level linguistic tasks, the model aggregates information over several layers until it reaches its highest performance. Often, the curves start steep, flatten out, and eventually drop again in the final layers.

Figure 1: Heatmaps illustrating our results for syntactic parent (P) and grandparent (GP) prediction (BERT-base, en, layers 1–12): Global metrics peak in middle layers. Local contributions are concentrated in early layers. (Darker shades indicate higher values.)

In this paper, we zoom in on models' *relative* information gains between one layer and the previous one. By this, we aim to turn the focus of probing away from information that is already present in the non-contextualized embedding layer and focus exclusively on information that needs context to be retrieved. We also aim to make the contribution of each layer to the model's overall linguistic capabilities explicit. We argue that under the hypothesis that information in language models is structured as in classical NLP pipelines (Tenney et al., 2019), the *depth* at which information emerges is as important as the *impact* of that information on model performance. When we think of language processing as a pipeline, we are primarily interested in where linguistic features emerge for the first time, not how long they are passed on to later layers.

To formalize information gains, we modify the conditional probing framework proposed by Hewitt et al. (2021). Their method explicitly quantifies information that is not already present in a baseline representation. We modify this method along two lines: First, we make it *local*, by conditioning on the respective previous layer instead of a global baseline. Second, we report the results as a *share* of total emergent information, across all layers of the network. This makes layer contributions comparable across tasks, where the overall performance may differ.

We demonstrate how to use our framework in practice, by applying it to two pairs of syntactic probing tasks. Each pair is formed by tasks that are structurally equal but for which we can reasonably assume a natural order in which the relevant linguistic representations emerge within the models. The first pair compares predicting part-of-speech (POS) tags that are the most frequent for a word form (MFTs) to predicting tags that are not (non-MFTs). Recent work has hypothesized that non-MFTs may be best represented in the deeper layers of a model such as BERT (Devlin et al., 2019): Each layer's contribution beyond the information already present in the uncontextualized layer is more significant for deeper layers, and therefore POS information could be found later in the model than previously assumed (Hewitt et al., 2021). The second pair of tasks compares predicting the position of a word's dependency head (the syntactic parent) to predicting the position of the head's head (the grandparent). Information for predicting grandparents has in previous work been found in deeper layers than information for predicting parents (Blevins et al., 2018). For each pair of tasks, we test where the relevant information emerges in the model, how the resulting pattern compares to global metrics, and if the metrics reflect the expected order of tasks. Our results show that while the expected hierarchy holds for the parent vs. grandparent task, information for non-MFTs emerges earlier in the model than for MFTs. This contradicts previous expectations. Also, results on seven independent monolingual BERT models in different languages show that the orderings and patterns we observe are not robust, which raises questions about their generality, and about the validity of probing at large.

2 Related Work

How linguistic information is distributed across the layers of a neural model is one of the central questions in the probing literature. The consensus is that there is a hierarchical ordering of tasks. Blevins et al. (2018) find a soft hierarchy of tasks when probing different layers of recurrent neural networks, from POS information being low in the hierarchy, to syntactic parents, grandparents, and great-grandparents. For their ELMo model, Peters et al. (2018) find that parts-of-speech are better predicted from the first hidden layer and word senses from the second. Tenney et al. (2019) probe BERT for a range of different NLP tasks and find that the layers that are the most predictive for each task are ordered like a classical pipeline: from parts-of-speech over syntactic dependencies, named entities and semantic roles to coreference.

How probing experiments should be designed and evaluated is subject to ongoing discussion. Some authors argue for simple classifiers (Alain and Bengio, 2017; Hewitt and Liang, 2019) to prevent the probes from learning the task and memorizing associations by themselves, while others make the case for more expressive models (Pimentel et al., 2020). While probes are most commonly evaluated using accuracy, recent work has proposed the use of alternative metrics that measure the effort of learning (Voita and Titov, 2020) or emphasize the performance early in the training (Talmor et al., 2020). Kunz and Kuhlmann (2021) propose to probe in an extrapolation setting, evaluating, among other setups, on the non-MFTs in diagnostic POS tagging experiments.

3 A Taxonomy of Metrics

We start by categorizing methods along three dimensions: The first one (as proposed by Hewitt et al. (2021)) concerns the relation of the baseline and the representation: how much more information can we extract from the representation than from the baseline (*baselined* probing), or how much information is extractable from the representation that does not overlap with information from the baseline (*conditional* probing).

The second dimension, proposed by us, concerns the type of information intended to be measured: information relative to a non-contextualized baseline (a *global* baseline), or information gain relative to the previous layer (a *local* baseline). The local setting challenges the view that a linguistic property's place in the model is the layer where most usable information for it can be extracted. Instead, we consider the layers where most usable information is *gained* relative to the previous layer to reflect the linguistic property's place within the model's hierarchy. We formulate and test the local correspondents of baselined and conditional probing in Sections 3.4 and 3.5.

Thirdly, we modify the local metrics so that they, in addition to the *absolute* reporting of the results, also support the reporting of the *relative* share that each layer contributes to the final performance. While absolute numbers convey more

information, relative numbers improve the comparability of results across tasks. We modify our local conditional metric from an absolute to a relative metric in Section 3.6.

3.1 General Setup

We consider a standard setup where we train a probe on a diagnostic task and evaluate it in terms of accuracy. More specifically, we use datasets $\mathcal{D} = \{(x_n, y_n)\}_n$ where each x_n is the representation of a neural language model at some specific layer, and y_n is the gold-standard label. (In our experiments, we use BERT.) By computing probe accuracy for different layers of the same model, we can compare layers in terms of how predictive they are with respect to the diagnostic task.

3.2 Global Baselined Probing (GBP)

In this common setup we measure the difference between the probe accuracy on a given layer l_i and the baseline layer l_0 – in BERT, this is the uncontextualized embedding layer. Thus we compute

$$\text{GBP}_i = \text{Acc}(l_i) - \text{Acc}(l_0) \tag{1}$$

As Hewitt et al. (2021) show, this can be interpreted as a difference between two quantities of $\mathcal{V}$*-information* (Xu et al., 2020), a theory of usable information under computational constraints. More specifically, GBP_i estimates the difference in $\mathcal{V}$-information between predicting the linguistic property under consideration from l_i and predicting it from layer l_0. This makes the difference in the probe's performance relative to the baseline explicit. The baselined information measures the amount of information gained over the baseline without making assumptions about the structural relation between l_0 and l_i.

3.3 Global Conditional Probing (GCP)

This setup has been proposed by Hewitt et al. (2021) with the intent to explicitly measure what information a layer l_i contributes *beyond* the information present in the baseline l_0. Practically, it entails computing the difference between the probe accuracy on the concatenation of l_i to l_0 and the baseline layer:

$$\text{GCP}_i = \text{Acc}([l_i; l_0]) - \text{Acc}(l_0) \tag{2}$$

In the framework of Hewitt et al. (2021), this measure is related to a conditional version of $\mathcal{V}$-information. More specifically, it estimates the conditional $\mathcal{V}$-information conditioned on prior information contained in the baseline.

3.4 Local Baselined Probing (LBP)

Analogously to global baselined probing, we may consider a local setup where the baseline is the previous layer l_{i-1}:

$$\text{LBP}_i = \text{Acc}(l_i) - \text{Acc}(l_{i-1}) \tag{3}$$

This quantity provides an estimate of how much $\mathcal{V}$-information is gained when taking the step from l_{i-1} to l_i. We posit that layers with high LBP values can be considered as layers where useful new information *emerges*. Intuitively, LBP measures the steepness of the slope, or the "jumps", in traditional accuracy curves across layers.

3.5 Local Conditional Probing (LCP)

To complete the picture, we propose to apply conditional probing to the local setting:

$$\text{LCP}_i = \text{Acc}([l_i; l_{i-1}]) - \text{Acc}(l_{i-1}) \tag{4}$$

The intention behind this metric is also to measure information gain with respect to l_{i-1}, but we account for exclusive information of l_{i-1} that is absent in l_i. Similar to Hewitt et al. (2021), we concatenate two layers and compare to scores on one of them. Our approach differs in that we do not compare to one static baseline layer (l_0) but dynamically to l_{i-1} to track the information gained across layers.

3.6 Emergent Information (EMI)

EMI (as well as EMI-BL in the next section) is designed to make layer contributions comparable across tasks that have different overall performances. To represent relative information gains, we calculate the LCP metric and divide it by the LCP summed up over all L layers. As we focus on *gains*, the metric will be zero whenever the result is negative (as nothing is gained). For the sum we also only consider layers where the LCP is positive.

$$\text{LCP}'_i = \max(0, \text{LCP}_i) \tag{5}$$

$$\text{EMI}_i = \frac{\text{LCP}'_i}{\sum_{k=1}^{L} \text{LCP}'_k} \tag{6}$$

We get relative gains that sum up to one. We interpret the results as the layer's contribution to the overall emergent information within the model.

3.7 EMI, Baselined Control (EMI-BL)

For ablation purposes we want to investigate the effect of the conditioning in EMI. Therefore we also employ a simplified version of EMI that uses LBP instead of LCP:

$$\text{LBP}'_i = \max(0, \text{LBP}_i) \tag{7}$$

$$\text{EMI-BL}_i = \frac{\text{LBP}'_i}{\sum_{k=1}^{L} \text{LBP}'_k} \tag{8}$$

The difference to EMI in Section 3.6 is the lack of control for information that was already present in the previous layer. We may underestimate the information gain as information may have been "forgotten" and replaced by new information when transitioning to the next layer.

4 Experiments

In our experiments, we apply the metrics defined in the previous section to study the performance of language models on two suitable probing tasks.

4.1 Probing Classifier

As our probe, we use a simple feed-forward network with 64 hidden units and ReLU activation, and train it for 10 epochs using the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 0.001. Our implementation uses PyTorch (Paszke et al., 2019). We calculate the results for all metrics based on the mean accuracy over 10 random seeds.

4.2 Language Representation Models

To test the robustness of our probing results and to explore if the syntactic information is localized in similar regions across models and languages, we include seven models in our analysis. Apart from English BERT (Devlin et al., 2019), we train probes on monolingual BERT models in Czech (Sido et al., 2021), Finnish (Virtanen et al., 2019), German (Chan et al., 2020), Hebrew (Seker et al., 2021), Swedish (Malmsten et al., 2020) and Turkish (Schweter, 2020). The languages are chosen to represent diverse families: Indo-European/Germanic (*de*, *en*, *sv*), Indo-European/Slavic (*cs*), Uralic (*fi*), Turkic (*tr*), and Afro-Asiatic/Semitic (*he*). All models are *base* models with 12 layers, and accessed via the Huggingface Transformers library (Wolf et al., 2020).[1]

[1]All code necessary to reproduce the results in this paper, along with full numerical results, is available here: https://github.com/jekunz/emergent_info

4.3 Data and Tasks

We consider two pairs of closely related syntactic probing tasks. The training data for these tasks is derived from 1,000 sentences randomly sampled from the Universal Dependencies treebank (Zeman et al., 2021).[2]

POS tagging We predict UPOS tags and evaluate on two sets, the most frequent tags for a word form (MFT) and tags that are not the most frequent for a word form (non-MFT). We assume that:

Hypothesis 1 *Models learn to predict non-MFTs in deeper layers than MFTs.*

Syntactic Ancestors Prediction We predict the relative linear position of a token's head (parent, P) and its head's head (grandparent, GP) in the syntactic dependency tree. For practical reasons we omit examples where the distance is larger than 15. For this task, our assumption is:

Hypothesis 2 *Models learn to predict grandparents in deeper layers than parents.*

Classically, we would also assume that the ancestors tasks come higher in the hierarchy than the part-of-speech tagging tasks, which would give us the following hierarchy of all tasks:

$$\text{MFT} < \text{non-MFT} < \text{P} < \text{GP}$$

However, as only the tasks in each pair are structurally equal, we will analyze each pair of tasks separately.

4.4 Ranking

To determine the hierarchical ordering of tasks within the models, we need to reduce the metrics across layers to a single comparable value. For that, we employ two strategies:

Max Layer For all metrics, we report the layer which maximizes the respective metric. When this layer is deeper for a task T than for a task T', we say that T is *higher* in the hierarchy induced by the model than T'.

Early Contributions For the EMI metric, we also report the contribution of layers 1, $1+2$ and $1+2+3$ to the overall gain. When this contribution is higher for a task T than for a task T', we say that T is *lower* in the hierarchy than T'.

	GBP (& Accuracy)		GCP		LBP & EMI-BL		LCP & EMI	
	MFT	¬MFT	MFT	¬MFT	MFT	¬MFT	MFT	¬MFT
cs	4	**6**	4	**6**	**3**	2	**3**	2
de	6	**11**	8	**10**	**4**	1	**4**	1
en	4	**7**	**10**	8	**2**	1	1	1
fi	3	**4**	**7**	5	1	1	1	1
he	3	**5**	**8**	7	**2**	1	**2**	1
sv	3	**5**	**11**	6	**2**	1	**2**	1
tr	2	**5**	3	**11**	2	2	2	2
avg	3.6	**6.1**	7.3	**7.6**	**2.3**	1.3	**2.1**	1.3

Table 1: Part-of-speech tagging tasks. The numbers give the layer of maximum score across metrics and languages. Bold marks the task (MFT or ¬MFT) that is higher in the hierarchy induced by the model.

	GBP (& Accuracy)		GCP		LBP & EMI-BL		LCP & EMI	
	P	GP	P	GP	P	GP	P	GP
cs	5	**8**	5	**8**	1	**2**	1	**2**
de	9	9	9	9	2	2	2	2
en	5	**6**	5	**7**	1	1	1	1
fi	5	5	5	5	2	**3**	2	**3**
he	5	5	**9**	5	**4**	3	**4**	2
sv	**7**	6	7	7	**2**	1	**2**	1
tr	**7**	8	7	**8**	3	3	3	3
avg	6.1	**6.7**	6.7	**7.0**	2.1	2.1	**2.1**	2.0

Table 2: Syntactic ancestors prediction tasks. The numbers give the layer of maximum score across metrics and languages. Bold marks the task (P or GP) that is higher in the hierarchy induced by the model.

5 Results

This section presents the results of our experiments. We have structured our presentation around the two ranking methods.

5.1 Max Layer

For each probing setup and language, we report that layer which maximizes the respective metric in Table 1 for the POS tagging task pair and Table 2 for the ancestors prediction tasks.

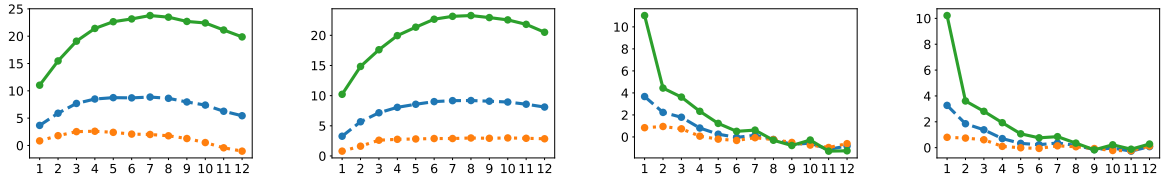
Global metrics Our results for the global metrics confirm the finding of Hewitt et al. (2021) that the layers that maximize conditional probing accuracy (GCP) are generally deeper than those that maximize baselined accuracy (GBP).

Zooming in on the distinction between most frequent and non-most frequent tags for the POS tasks, however, exhibits an unexpected behavior: Hewitt et al. (2021) suggest that for the non-MFTs, GCP should be higher than GBP in deeper layers, and the other way round for MFTs. Here we find that in 4 out of 7 models, the layer with the highest GCP value on non-MFTs *precedes* the layer with the highest value for MFTs. However, the average over the models is higher for non-MFTs due to the large margin between the layers in the *tr* model. The highest scores of GBP on non-MFTs are consistently in deeper layers than those for MFTs. The exact layer in which the maximum scores are however varies greatly between models: for the MFTs, it ranges between 2 (*tr*) and 6 (*de*) and for the non-MFTs between 4 (*he*) and 11 (*de*).

GCP differentiates less than GBP, with a margin of 2.5 versus 0.3 (POS tagging) and 0.6 versus 0.3 layers (ancestors prediction) difference between the lower-level and the higher-level task.

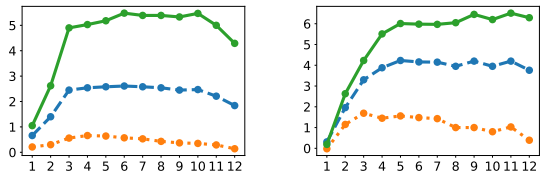
Looking at the full plot, rather than just the maximal layer, we observe some variety across metrics and languages. The example plots for *en* BERT in Figure 2 (a–b) are in line with the general trend: GCP peaks in deeper layers than GBP, but this is not explained by the non-MFTs, as their curve drops steeper with increasing layer index than the curve for the MFTs. This observation holds for most BERT models we used, except for *cs* and *tr* where the scores on MFTs drop more in deeper layers than those for non-MFTs (see Figure 3).

[2]The treebanks for each language are: *cs*: PDT, *de*: GSD, *en*: EWT, *fi*: TDT, *he*: HTB, *sv*: Talbanken, *tr*: Kenet. Lic: CC BY-SA 4.0 (*de*, *en*, *fi*, *sv*, *tr*) / CC-BY-NC-SA 3.0 (*cs*, *he*).



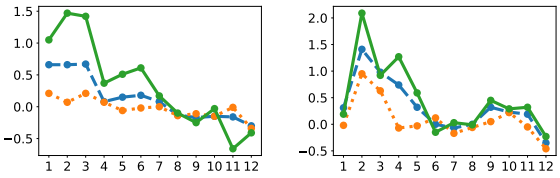
(a) GBP (b) GCP (c) LBP (d) LCP

Figure 2: Part-of-speech tagging, global (a–b) and local (c–d) metrics on the English data. Solid green line: non-MFTs, dotted orange: MFTs, dashed blue: full development set (all tags).



(a) GCP Czech (b) GCP Turkish

Figure 3: As opposed to *en* BERT and the other four models, for *cs* and *tr*, the scores on MFTs in GCP drop more over the layers than those for non-MFTs.



(a) LCP Czech (b) LCP Turkish

Figure 4: For *cs* and *tr* BERT, the LCP plots exhibit a pattern that deviates from that we observe for *en* BERT: They do not decrease steadily.

Local metrics Turning to the highest local information gain in Table 1 and 2, both LBP and LCP show the biggest gains in the very first layers. This shows that, while global metrics locate the overall information peaks somewhere in the middle of the model, only little new information actually emerges at that point.

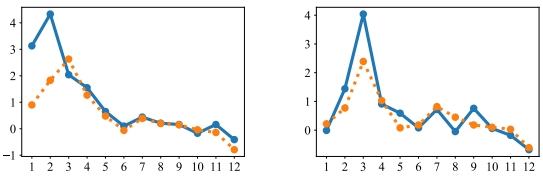
The differences in the empirical results between LBP and LCP are small; specifically accounting for information that is absent in the previous layer does not result in a different pattern than the one we obtain when using the baselines metric. This leaves the choice between the two metrics to theoretical or practical preferences.

The example curves for English BERT in Figure 2 (c–d) show a typical pattern for the drop across layers in the part-of-speech tagging tasks. The layer of highest information gain appears to be the layer where contextual information is added first. After this layer, the plots decrease more slowly. Most languages follow this pattern, with the notable exception of *cs* and *tr* that do not show a steady decrease but go up first (see Figure 4). For the ancestors tasks, the peak is often shifted to the second or third layer, probably reflecting the higher-level nature of those tasks compared to the POS tasks. Figure 5 shows two examples, *fi* BERT with and *tr* BERT without a clear hierarchy of the parents versus grandparents task.

We observe that for the local metrics, the supposedly higher-level tasks do *not* have their highest gains in later layers than the lower-level tasks. On the contrary, the non-MFTs in LBP and LCP have their average max layer at 1.3, while for MFTs, where the accuracy starts off much higher, and information gains are generally smaller, the corresponding values are 2.3 and 2.1. For the ancestors tasks, there is on average no difference between the tasks. Hence, Hypotheses 1 and 2 are not confirmed for the local metrics in the max layer rankings.

5.2 Early Contributions

As Hypotheses 1 and 2 about the order of tasks in the model's hierarchy were not confirmed when looking at the layer with the maximal score, we compare more expressive metrics from the emergent information family in Tables 1 and 2.



(a) LCP Finnish (b) LCP Turkish

Figure 5: Ancestors prediction, LCP: *fi* shows a later peak for grandparents (orange), while *tr* BERT's curves show a similar pattern for both tasks.

	Layer 1		Layer 1 + 2		Layer 1 + 2 + 3	
	MFT	¬MFT	MFT	¬MFT	MFT	¬MFT
cs	37.49	**18.74**	49.99	**44.99**	87.49	**70.35**
de	**0.00**	34.18	**23.85**	62.32	**46.78**	74.94
en	**31.00**	46.24	**59.68**	62.57	83.33	**75.29**
fi	**44.07**	55.28	**58.12**	71.87	**76.30**	84.31
he	39.31	**37.93**	89.74	**74.94**	97.15	**89.69**
sv	**12.86**	52.23	**48.53**	78.70	**58.47**	89.00
tr	**0.00**	3.08	48.22	**37.07**	80.20	**52.03**
avg	**23.53**	35.38	**54.01**	61.78	**75.67**	76.51

Table 3: Part-of-speech tagging tasks: Contribution of layer 1, $1 + 2$ and $1 + 2 + 3$ to the overall performance of the probe. Bold marks the task (MFT or ¬MFT) that is is higher in the hierarchy induced by the model (smaller contribution of the lower layers).

	Layer 1		Layer 1 + 2		Layer 1 + 2 + 3	
	P	GP	P	GP	P	GP
cs	46.78	**14.90**	73.39	**68.02**	**79.10**	81.30
de	8.68	**3.94**	50.06	**35.49**	68.68	**51.83**
en	37.47	**36.86**	55.15	**52.79**	78.89	**56.81**
fi	24.51	**11.46**	58.41	**34.77**	74.39	**68.28**
he	0.00	0.00	**16.90**	25.74	**27.46**	51.48
sv	**21.14**	39.31	60.33	**49.60**	77.65	**66.61**
tr	**0.00**	3.52	16.74	**15.84**	63.72	**54.08**
avg	19.79	**15.71**	47.28	**40.32**	67.12	**61.48**

Table 4: Syntactic ancestors prediction tasks: Contribution of layer 1, $1 + 2$ and $1 + 2 + 3$ to the overall performance of the probe. Bold marks the task (P or GP) that is is higher in the hierarchy induced by the model (smaller contribution of the lower layers).

For the MFT tasks, we can confirm the finding that when looking at the share of emergent information from the very first layers, it is on average higher for the non-MFTs for all three groupings of layers, indicating that non-MFTs would be lower in the hierarchy that MFTs. The difference however fades out, from 11.85 percentage points difference for layer 1 to only 0.84 points for layer $1 + 2 + 3$. For the ancestors tasks, the hierarchy continues to be as expected both on average and in the vast majority of model-grouping combinations. For both tasks we note that no layer grouping shows consistent results across all seven models, indicating a low robustness of the results.

Effects of Conditioning The average difference between emergent information with or without conditioning on the previous layer is 5.58 (POS) and 3.93 percentage points, which can in some scenarios be considered as minor. For all values we examined in this section, the average difference between tasks is 16.93 (POS) and 11.14 (ancestors) points, making a trend change unlikely and a simplification towards a baselined setup justifiable. Less distinct tasks and their theoretical advantages may however suggest the inclusion of the conditional accuracy.

6 Discussion

Where is a feature located? The results for the different metrics show how it depends on the perspective which place within the model we assign to a linguistic property. While the most overall information is located in the middle layers, it is the early layers that maximize the local metrics by a huge margin, meaning that this is the place in the model where most information either emerges or becomes accessible.

The complementary use of both families of metrics, local and global, gives us a more holistic picture of how information is structured within the model, as we argue that it it is not clear if a hierarchy of tasks within the model should be determined by where most information is added or where most information is accessible overall. An intuition supporting the former is the comparison to a human-made pipeline model, where tasks would naturally be placed where the information required for them is added – for instance, where the POS tagger is located, adding POS tags to the set of features. In these pipelines, there is no notion of how long information will be passed to higher-level tasks.

Expected hierarchies do not generally hold We see that when probing for MFTs versus non-MFTs, the perceived natural hierarchy of tasks (Hypothesis 1) does not hold for the local metrics, neither in a coarse max layer analysis nor in the more fine-grained early contribution setting, as shown in Sections 5.1 and 5.2. In the plots of the global metrics in Section 5.1 we see that non-MFTs often show both steeper gains in the beginning and more pronounced losses in the later layers, indicating that it is more specialized contextual information that the non-MFTs require, but that information does not appear to emerge later than in the model that for MFTs. We conclude that observations of a clear hierarchy of tasks depend on the focus on most usable overall information: They are already weaker in the global conditional setup, and are in one of two cases contradicted by local metrics.

The massive dominance of layer 1 and 2 in all local metrics especially for the non-MFTs but even for the parents and the grandparents tasks raises questions about how high-level the information that contributes the most to the overall performance on the probing task actually is. As the very first possibility of accessing contextual information already presents the heaviest boost, the features that are most crucial to solve the task appear to be surprisingly shallow.

Probing results are not robust An interesting point we noted across all metrics, but more distinctly for all local metrics as well as global conditional probing, is that the results are not stable across BERT models in different languages. We do not consider the possibility to relate the different distribution of information across models to linguistic properties of the languages as we believe that this is impossible with the relatively small set of non-parallel models we analyze. Apart from the language, they differ in several variables: most importantly, the data they are trained on, but some also in training details. However, we see it as an exciting path for future research to explore what causes a model to structure information in certain ways, and if this has implication on the model's performance on downstream tasks or robustness.

Relevance As the differences between measuring emergent information with and without conditioning on the previous layer are relatively small, one could suggest that the information is present in a similar form in ordinary accuracy plots across lay-

ers: The slope of the curve can be used to estimate it. But in practice, such interpretations do not appear to be obvious: Even though they can, as we showed, shed a different light at probing results when made explicit, previous work exclusively focused on the point of highest overall information. A discussion on the relevance of the parts of the network where information emerges has been absent from the literature.

Apart from that, we argue that conditional probing has strong theoretical advantages, as it explicitly accounts for information in the baseline representation, and that this makes basing the emergent information metrics on it favorable.

Limitations The metrics we propose are designed with the expectation that gains are successive. However, Transformer models can propagate information via residual connections and thereby let information "skip" layers. If this resulted in pronounced oscillations of the information within the model, it would weaken the meaningfulness of the results of all local metrics.

The emergent information metrics have no account for loss of information over the layers of the network. A related family of metrics that explicitly models this would be the application of the local metrics to the layers in inverse order.

All metrics in this paper are based on probe accuracy. However, our setups can be easily adapted to other metrics which have been shown to be more robust towards design choices regarding the classifier, such as minimum description length (Voita and Titov, 2020), or metrics that reward fast learning (Yogatama et al., 2019; Talmor et al., 2020).

7 Conclusion

We have collected and suggested metrics that model the information distribution in a model's layers from different perspectives: globally and locally, with or without conditioning on the baseline, and looking at absolute and relative gains of information. We used them on two pairs of probing tasks. First, we tested whether information for POS tags that are not the most frequent for a word is found in deeper layers than general POS information and found that while this is the case for overall information measured by global metrics, local metrics highlight that the most significant gains consistently happen in the very first layers *in particular* for the non-most frequent tags. For second task of predicting the syntactic parents versus grandparents of a token, however, the expected hierarchy in the model holds in the local setup at least in more fine-grained relative metrics. These mixed results emphasize the additional insights that zooming in to local information gains can give us into the model, the task, and the probing methodology.

Probing experiments on seven monolingual BERT models in different languages show that the metrics' behavior varies between models. While it is currently not feasible to relate the differences to specific properties of the models such as the language or the domain of the training data, a controlled training of parallel models where the additional variables are controlled for may enable such a comparison and is an insightful direction for future work.

References

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net.

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

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, Melbourne, Australia. Association for Computational Linguistics.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher Manning. 2021. Conditional probing: measuring usable information beyond a baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1626–1639, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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 (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

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

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Jenny Kunz and Marco Kuhlmann. 2021. Test harder than you train: Probing with extrapolation splits. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 15–25, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with words at the National Library of Sweden – Making a Swedish BERT.

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 H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

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

Stefan Schweter. 2020. BERTurk - BERT models for Turkish. *Zenodo*.

Amit Seker, Elron Bandel, Dan Bareket, Idan Brusilovsky, Refael Shaked Greenfeld, and Reut Tsarfaty. 2021. AlephBERT: A Hebrew large pre-trained language model to start-off your Hebrew NLP application with. *arXiv preprint arXiv:2104.04052*.

Jakub Sido, Ondřej Pražák, Pavel Přibáň, Jan Pašek, Michal Seják, and Miloslav Konopík. 2021. Czert – Czech BERT-like model for language representation. *arXiv preprint arXiv:2103.13031*.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. oLMpics-on what language model pre-training captures. *Transactions of the Association for Computational Linguistics*, 8:743–758.

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, Florence, Italy. Association for Computational Linguistics.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Finnish. *arXiv preprint arXiv:1912.07076*.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 Rush. 2020. 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, Online. Association for Computational Linguistics.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. 2020. A theory of usable information under computational constraints. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Dani Yogatama, Cyprien de Masson d'Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, et al. 2019. Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373*.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy

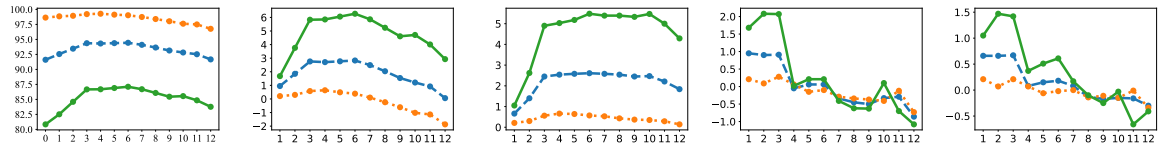
Ajede, Gabrielė Aleksandravičiūtė, Ika Alfina, Lene Antonsen, Katya Aplonova, Angelina Aquino, Carolina Aragon, Maria Jesus Aranzabe, Bilge Nas Arıcan, Hórunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Deniz Baran Aslan, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Starkaður Barkarson, Rodolfo Basile, Victoria Basmov, et al. 2021. Universal dependencies 2.9. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

A All results

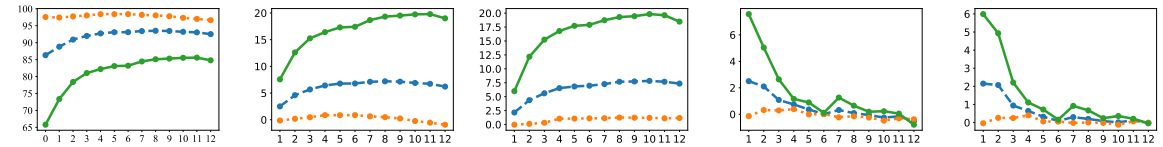
For completeness, we present all plots across models and metrics as supplementary material in figure 6 for the POS tasks and in figure 7 for the ancestors tasks. A brief summary of the material is provided in the following paragraphs.

Global Metrics. The accuracy shows the same highs and lows as the GBP setup, where the static l_0 baseline is subtracted from the accuracy. Compared to GBP, the results in the GCP setup are slightly shifted to later layers. For the POS tasks, the peak is in the early middle layers, with the ¬MFTs peaking a few layers later, indicating the need for more contextual information. Across models we see a large variation, most extremely visible in *de*, where the scores increase until layer 11 for the MFTs, and tr, where the drop for the MFTs is more distinct than for other models. *fi* and *he* have a distinct peak for the ¬MFTs in layer 4, then a decrease, and then stabilize. The ancestors tasks often peak in the early middle layers as well, with *de* being shifted to notably later layers, and *tr* being relatively stable across all layers except the very first and last layers. A later peak of the grandparent prediction compared to the parent prediction is vaguely perceptible in most plots, most prominently in the *en* model.

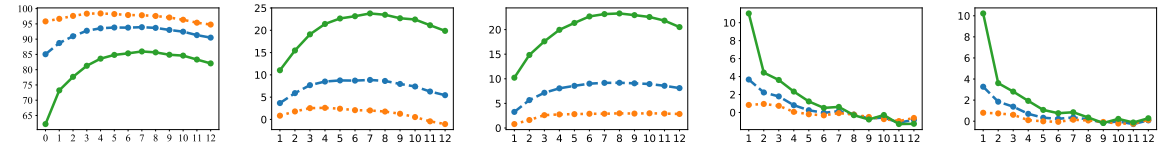
Local Metrics. The metrics that measure the local information gain have the most consistent pattern for the ¬MFTs, with most information generally added in the very first layer. The pattern of the curves appears to asymptotically approximate 0. There are however two exceptions: the *cs*, but most distinctly the *tr* model that gains relatively little in the first layer and makes its biggest jump in the second layer. We also observe in the accuracy curve of these two models that the increase in the beginning is less steep. In the ancestors tasks, the highest layer is slightly later on average, often in later 2 or 3 (2.1 on average). In some models, such as *cs* and *fi*, we observe a later peak for grandparents than for parents, while for *de*, and *se*, it even is the other way round. This underlines the lacking robustness of our probing results across models in different languages that are particularly prominent for the local metrics. In all of the models we observe little difference in the empirical results and patterns of LBP and LCP, confirming our observations in Section 6 that the choice between them can be either arbitrary or based on theoretical preferences.



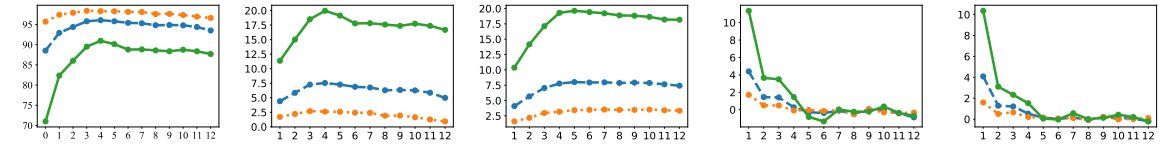
(1) Accuracy Czech (2) GBP Czech (3) GCP Czech (4) LBP Czech (5) LCP Czech



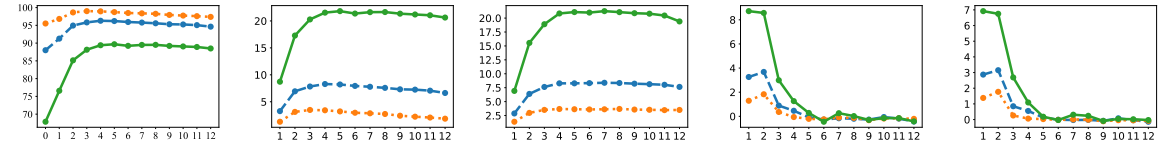
(6) Accuracy German (7) GBP German (8) GCP German (9) LBP German (10) LCP German



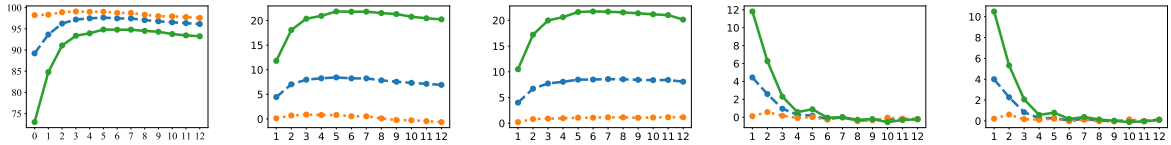
(11) Accuracy English (12) GBP English (13) GCP English (14) LBP English (15) LCP English



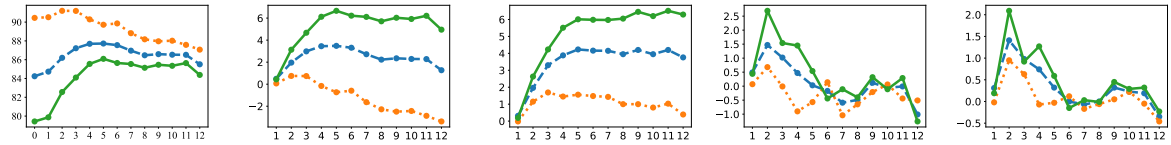
(16) Accuracy Finnish (17) GBP Finnish (18) GCP Finnish (19) LBP Finnish (20) LCP Finnish



(21) Accuracy Hebrew (22) GBP Hebrew (23) GCP Hebrew (24) LBP Hebrew (25) LCP Hebrew



(26) Accuracy Swedish (27) GBP Swedish (28) GCP Swedish (29) LBP Swedish (30) LCP Swedish



(31) Accuracy Turkish (32) GBP Turkish (33) GCP Turkish (34) LBP Turkish (35) LCP Turkish

Figure 6: POS Experiments: Plots for all language/metric combinations. Orange: MFT; green: ¬MFT; blue: all.

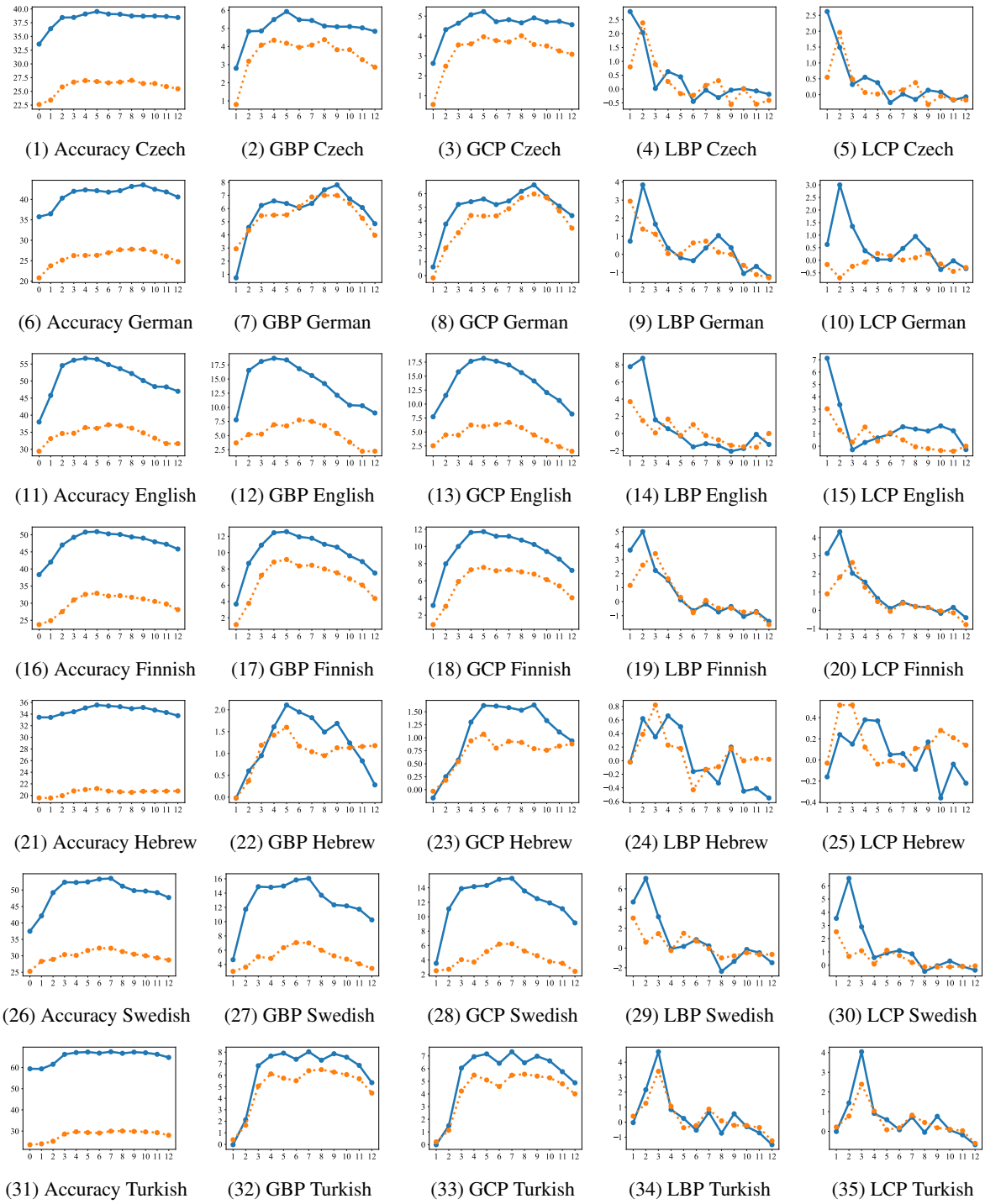


Figure 7: Ancestors Experiments: Plots for all language/metric combinations. Blue: P; orange: GP.