ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет «Высшая школа экономики»

Факультет гуманитарных наук
Образовательная программа
Фундаментальная и компьютерная лингвистика

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

На тему «Пробинг усвоения языка в процессе обучения языковых моделей»

Тема на английском “Probing language transformer-based models for language acquisition”

Студентка 4 курса
группы №182
Волошина Екатерина Юрьевна

Научный руководитель
Толдова Светлана Юрьевна
к.ф.н., доцент Школы лингвистики ФГН

Научный консультант
Сериков Олег Алексеевич
приглашенный преподаватель
Школы лингвистики ФГН

Москва, 2022 г.

Abstract

Modern state-of-the-art language models have shown high results not only on text generation tasks but on natural language understanding tasks. Therefore, in recent years, the models were tested on different probing tasks to examine their language knowledge. However, few researchers explored the very process of models' language acquisition. Nevertheless, the analysis of language acquisition during training can show what types of language data are easier to acquire and how the process of acquisition depends on model architecture or model sizes.

For the natural language processing field, this research could shed light on the model parameters that help to acquire the language faster. From the linguistic point of view, such analysis could find correlations between language structures of different levels. If the language functions as a level system, models will acquire it in the order of language levels.

In this thesis, a new methodology for probing is presented. We perform several experiments on two English models checkpoints (MultiBERT (Sellam et al., 2021) and T5 (Raffel et al., 2019)) and then compare their results. We find that linguistic information is acquired during early learning steps, especially morphology and syntax. Moreover, the model architecture seems not to influence the language acquisition process. We also experiment with model sizes and reveal that the hidden size is the most essential factor for model language acquisition.

Contents

1. Introduction

Modern deep learning models have achieved significant results in the field of language modeling and text generation (Krause et al., 2019; Niu et al., 2020). Therefore, language models (LMs) are often used in linguistic research to find systematic similarities in the language data. Performance of the state-of-the-art models, such as Transformer-based ones (Vaswani et al., 2017), on linguistic tasks show that they have learned measurable language structures during the training process.

Consequently, it is interesting to explore how the LMs acquire the language during their training process. Doubtlessly, the language models learn the language in a different way than humans do. However, a closer look at the models acquisition can shed light on the correlations between different levels of language structure, for example, to reveal if it is necessary to acquire syntax first to achieve high results on discourse tasks. According to the theory of language levels, any language consists of five layers, and they function in the following order: phonology, morphology, syntax, semantics, and discourse.

In this thesis, we research if the process of models' language acquisition follow the theory of language levels. Moreover, we will study the correlation between the acquisition process and different model architectures, model sizes, and different linguistic tasks. Linguistic tasks are meant to represent three levels of language structure: morphology, syntax, and discourse. We leave behind tasks on phonology, since we work with written texts, and semantics, which was often researched previously.

If the language models distinguish between different levels, they should show differences in the process of acquisition on these tasks. In terms of model sizes and architectures, it is expected that if the order of acquisition is the feature of a language, and not a model, models of different sizes and architectures should show the same results on different probing tasks. In other words, we pose the following questions: (i) is there enough empirical evidence to say that language models acquire the language? (ii) when do models acquire language structure, in other words, how many steps are needed for language acquisition? (iii) does language acquisition process depend on the level of grammar that a feature belongs to? and (iv) which parameters of models influence the language acquisition process?

The rest of the paper is structured in the following way: in the next section, we describe essential theory for our research, such as transformers and probing methodology. In the third

section, we give a brief overview of recent works on probing methods and recent research on LMs acquisition patterns. In the fourth section, we will describe models, probing tasks, probing methods and measures of acquisition level used in the research. In the fifth section, we report the results that we have achieved. The last section is dedicated to the discussion of our results and their value to the field of natural language processing.

2. Theoretical background

2.1. Transformer architecture

In this thesis, we work with transformer-based models as they are modern state-of-the-art neural networks. Transformer architecture was introduced in Vaswani et al. (2017). Transformers were made for sequential data, and the original architecture included an encoder and a decoder (see Figure 1). The encoder maps input sentences to vector representations, and decoder generates output from these vector representations. Transformers were originally proposed for machine translation, therefore, the encoder received sentences in source language and the decoder generated sentences in target language.

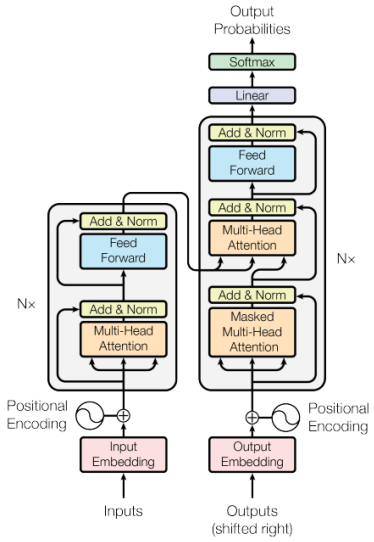

Figure 1: Transformer architecture from Vaswani et al. (2017)

The encoder consists of six layers, each one includes sublayers, a multi-head attention sublayer, which will be described later, and a linear sublayer of hidden size n. After each sublayer, an embedding goes through normalisation by layer. The decoder also consists of six layers, but each layer includes one more part: multi-head attention over a vector representation from an encoder stack. Both the encoder and the decoder include encoding of a token's position in a sentence.

The main innovation of this architecture lies in multi-head attention mechanism. As an input, an attention layer has three vector projections: query, key, and value. Key and value are linear transformations of a vector representation of an encoding token, and query is a linear

transformation of a previous token, except for self-attention, where query is also taken from the same token.

First, query is multiplied with all keys. Their matrix product is then scaled and masked if that's required. After applying *softmax* function to the product of query and keys, it is multiplied with values. The output of this layer is the weighted sum of the resulting values.

The attention layer is applied to several projections in parallel, so they could learn different information simultaneously. First calculated, all heads are then concatenated.

The encoder and the decoder have different attention layers. The encoder stack contains self-attention layers. In these layers all queries, keys and values come from the previous layer of the encoder. Each position has access to all positions from a previous layer. The idea of a self-attention layer is to reflect the context by replacing each element by weighted average of the other elements in the sequence. In the decoder, a multi-head attention sublayer masks forthcoming context with zeros, in other words, it "hides" upcoming tokens, so, predictions would only depend on the previous context.

BERT introduced in Devlin et al. (2018) is based on Transformer encoder architecture with bidirectional self-attention heads. BERT was trained on two tasks: Masked Language Modelling (MLM) and Next Sentence Prediction (NSP). In the first task, the data generator chooses 15% of tokens to mask, and then in 80% of examples it masks a chosen word with a special token [MASK] and in 20% with a different word. A model is given a sentence with a masked token, so it has to fill the gap with the right token. The second task was meant to teach a model to catch relationship between two sentences. The training data included 50% of pairs of sentences where one sentence goes after another in an original text and 50% of pairs of random sentences. The model's aim is to predict if the second sentence is truly the next sentence to the first one. The changes in the architecture and training approach let BERT to show the best results at that time on GLUE benchmark (A. Wang et al., 2018).

Another architecture based on original Transformer architecture was Text-to-Text Transfer Transformer (T5) presented in Raffel et al. (2019). T5 is an encoder-decoder model and it follows closely the original Transformer architecture described above, except that T5 has a different position embedding scheme and does not have Layer Normalisation. T5 was trained on several text-to-text supervised and unsupervised tasks. For supervised tasks, an input includes a prompt (for example, "summarize:" or "translate English to German:") and a source

sequence, and the output of a model should be generated according to a given prompt. For unsupervised tasks, the model gets a sequence with several masked tokens, and it has to reconstruct all masked tokens and mask all unmasked tokens. Such training setup made T5 to beat state-of-the-art models at the moment of T5 realisation on GLUE benchmark.

2.2. Probing methodology

Probing tasks were introduced in Conneau et al. (2018). As the authors state, "a probing task is a classification problem that focuses on simple linguistic properties of sentences". Probing tasks use an encoded vector representation of sentences to predict some linguistic features contained in these sentences. Probing methodology shows in what extent models behaviour can be explained with linguistic theory. As Saphra (2021) mentions, the main assumption behind probing is that models require knowledge that they are tested on for better performance on natural language generation and other tasks that a model can be used for. The probing methods are believed to help improve model performance and make it comparable with human performance (McCoy et al., 2019).

The basic probing method is to take a sentence vector from a language model that researchers want to probe and then give it as an input to a classifier, for example, logistic regression. The task of this diagnostic classifier is to put a label of linguistic feature to each sentence vectors. For example, for a number task, a diagnostic classifier has to assign a label of *plural* or *singular* on a sentence vector. Probing tasks are meant to address a simple problem, such as tense or a noun's number, to exclude possible bias or interpretability problems.

Diagnostic classifiers are not the only probing method. Belinkov et al. (2020) classifies existing methods as *structural*, which include diagnostic classifiers, and *behavioural*. Unlike structural probes, behavourial probes do not require any classifier on top of vector representations from a model. An example of behavourial probes is a masking task when a language model that is probed has to fill in a masked token, for example, to put a right verb form in a sentence.

3. Related work

It is a commonplace in linguistics to represent language as a system of several levels: phonology, morphology, syntax, semantics, and discourse. This theory also implies that levels operate in a specific order, for instance, phonology operates first and creates elements that morphology unites into morphemes and so forth. However, some researchers (Caha, 2009; Embick and Noyer, 2007) doubt that a language functions as a level system. They suggest that morphology and syntax operate at the same time and as one system.

In spite of theoretical debate on existence of a level system, it is often used as a ground for studies in different fields of linguistics, such as sociolinguistics (Van Bezooijen and Gooskens, 1999), clinical linguistics (Crystal, 1987) or computatinal linguistics (for example, in Conneau et al. (2018) tasks are classified as 'surface information', 'syntactic information', and 'semantic information'). In language acquisition, tasks are usually constructed following the hierarchy of language levels .

The main purpose of computational approaches to language acquisition is to find limitations of the knowledge that can be learnt without any special abilities, an example of which is a biologically evolved bias claimed in Universal Grammar (Chomsky, 1965). The belief that children acquire language due to biological biases is called *linguistic nativism*. Some researchers use computational models as a proof for the existence of such biological bias because models do not acquire language on the level of humans. However, as Clark and Lappin (2010) claim, "computational learning theory does not motivate strong linguistic nativism, nor is it irrelevant to the task of understanding language acquisition". The authors mention that none of existing models copies the human acquisition process, however, models can be seen as computational idealisations that clarify the class of learnable features of natural language.

Computation research on language acquisition has two different directions: computational models of child language development, known as statistical language acquisition, and an insight into pre-trained language models' training process. Statistical language acquisition uses simple neural networks to simulate child acquisition. For example, Lewis and Elman (2001) use an RNN model to distinguish sentences with inversion and without it. Prefors et al. (2006) use Bayesian learning to research the structure dependency problem: the model has to select between a finite state grammar and context-free grammar.

The tasks used in statistical language acquisition are similar to probing tasks. Unlike

models used in statistical language acquisition, models used in probing are usually pre-trained and have shown significant results on other tasks, such as text generation.

Most of the probing studies explore the limitations of pre-trained models. Models are usually evaluated on different tasks across all layers. However, other parameters were explored as well, such as the number of tokens needed to acquire different features (Zhang et al., 2020) or the number of iterations needed to learn basic linguistic information.

The first work on probing of neural networks across time was carried by Saphra and Lopez (2018). The authors showed that first, a LSTM model (Hochreiter and Schmidhuber, 1997) acquires syntactic and semantic features and later information structure. Chiang et al. (2020) looked at the training process of ALBERT (Lan et al., 2019) and concluded that semantic and syntactic information is acquired during the early steps while accuracy on world knowledge fluctuates during the training. L. Z. Liu et al. (2021) showed similar results on RoBERTa (Y. Liu et al., 2019): the model shows good results on linguistic probing tasks starting from early stages, and later it learns factual and common sense knowledge. (Blevins et al., 2022) studied training dynamics of multilingual models, they reveal that while linguistic information is acquired early, transfer learning abilities are evolving during the entire training process. Choshen et al. (2022) examined the trajectories of models' language acquisition, and they find no impact of architecture or a model size on training trajectories.

It is worth noting that high results of models on probing tasks may be not only due to the generalisation abilities of models but also because of the structure of data. To make probing results more reliable in terms of reflection of model knowledge, new approaches to probing were proposed, such as Minimum Description Length (MDL) (Voita and Titov, 2020) that is based on information theory. This method shows how much of language information is encoded in embeddings.

4. Methods

4.1. Models

We calculate the accuracy of two transformer models on 12 probing tasks. As we want to know how universal patterns of language acquisition in models are, we experiment with two different transformer architectures: BERT and T5.

While BERT has only encoder layers, T5 includes both encoder and decoder layers. Therefore, embeddings from BERT come from the encoder, and T5 embeddings are calculated after going through the encoder and the decoder.

For this work, we use already published models with available checkpoints. It means that they were trained on different data and computational powers. The further comparison of two models is indicative only.

MultiBERT (Sellam et al., 2021) is based on BERT-base-uncased architecture, and it is the model of the same size (12 layers and embedding size 768). Unlike the original BERT (Devlin et al., 2018), it was trained with 25 different seeds. The authors also released checkpoints of the first five models. We use the model with seed 0 in our experiments.

T5-small[1] was trained by DeepPavlov laboratory. It is based on the Hugging Face implementation of T5 (Raffel et al., 2019), so it consists of 6 layers with 512 embedding size.

As a baseline, we use the method described in Hewitt and Liang (2019). We train logistic regression on top of embeddings of models mentioned above with shuffled class labels.

Since we want to study not only when and how well models acquire linguistic features but also models of what size can learn language on a good level of performance, we conduct several experiments on model sizes to explore the parameters of models that might be essential for language acquisition. We regard number of layers, embedding size and number of attention heads to be crucial. Therefore we train four models:

1. A tiny model: the hidden size of 128, 2 layers, and 2 attention heads;
2. A model with increased number of attention heads: the hidden size of 128, 2 layers, and 4 attention heads;
3. A model with increased hidden size: the hidden size of 256, 2 layers, and 2 attention heads;

[1] https://github.com/yurakuratov/t5-experiments

Task	Acceptable sentence	Unacceptable sentence
Transitive	*The pedestrians question some people.*	*The pedestrians wave some people.*
Passive	*Tracy isn't fired by Jodi's daughter.*	*Tracy isn't muttered by Jodi's daughter.*
Principle A c command	*This lady who is healing Charles wasn't hiding herself.*	*This lady who is healing Charles wasn't hiding himself.*
Adjunct Island	*Who does John leave while alarming Beverly?*	*Who does John leave Beverly while alarming?*

Table 1: BLiMP Minimal pairs examples

4. A model with increased number of layers: the hidden size of 128, 4 layers, and 2 attention heads.

If any of these models show a significantly different result on any group of tasks, this parameter causes a better acquisition process, which will help to train models in a different way, stimulating them to acquire the language. If all the models show similar results, it will mean that different sizes of models do not correlate with the acquisition process, therefore, it depends on language features rather than on model parameters.

We train models with the same computational resources and data corpus, which included Wikipedia articles limited to 10,000,000 tokens. We choose this threshold as an optimal one, since Zhang et al. (2020) proved that models can acquire basic linguistic information from this amount of data. We use the same seed as MultiBERT which was described above to make our results more comparable.

4.2. Probing tasks

We use probing tasks that come from several datasets published earlier: SentEval (Conneau et al., 2018), Morph Call (Mikhailov et al., 2021), DisSent (Nie et al., 2019), DiscoEval (Chen et al., 2019), and BLiMP (Warstadt et al., 2020).

The datasets from Benchmark of Linguistic Minimal Pairs (BLiMP) have a structure different from other tasks: every task includes pairs with minimal differences to illustrate one of the grammatical features of English. One sentence of the pair is grammatical, whereas another one is unacceptable. We choose four BLiMP tasks for our experiments: transitive and passive verbs, Principle A of C command, and Island effects. For the first two tasks, pairs have different verbs, where only one verb is transitive or can be used in a passive form. These tasks are categorized as morphological (see Table 1).

The next two tasks reflect syntactic effects on English. The Principle A task shows the use of reflexives. According to Chomsky (1981), a reflexive should have a local antecedent,

and if it does not, the sentence is ungrammatical.

The task on Island effects tests a model's sensibility to syntactic order. An island is a structure from which a word cannot be moved (Ross, 1967). Thus, the sentence is unacceptable if a word is placed out of an island.

The tasks from other datasets are summarized below (see examples in Table 2 and Figure 7 illustrating dataset balance):

- **Subject number** (SentEval): this task is supposed to show how models acquire morphology. It is a binary classification task with labels NNS and NN (plural and singular number, respectively). The classifier should decide on a sentence class based on the number of sentence subjects.
- **Person** (Morph Call): this task is also morphological. It is a binary classification with labels 0 and 1, which signifies if a subject has a person marker or not.
- **Tree depth** (SentEval): this task contains six classes, each of which stand for a depth of the syntactic tree of a given sentence. Hence, this task is meant to show the level of syntax acquisition.
- **Top constituents** (SentEval): this multiclass task belongs to the group of syntactic tasks. The aim is to choose a class that includes constituents located right below the sentence (S) node.
- **Connectors** (DisSent): this dataset includes pairs of sentences originally connected with one of 5 prepositions, and the task is to choose the omitted preposition. It is supposed to show how models catch discourse relations.
- **Sentence position** (DiscoEval): this dataset contains sequences of 5 sentences, and the first sentence is placed in the wrong place. Therefore, the aim is to detect the original position of these sentences. This task is also meant to show the models' accuracy on discourse.
- **Penn Discourse Treebank** (DiscoEval): the task is based on Penn Discourse Treebank annotation (Marcus et al., 1994). The aim is to choose the right discourse relation between two discourse items from Penn Treebank.

- **Discourse coherence** (DiscoEval): this task is a binary classification with classes 1 and 0. Class 1 means that the given paragraph is coherent, and class 0 should be assigned to paragraphs with shuffled sentences.

4.3. Probing methods

Sentence embedding classification: Token embeddings from transformer models are turned into a sentence embedding via mean pooling. A classifier model is used to classify embedded sentences. This method is used with tasks from SentEval and Morph Call.

Positional sentence classification: For the Sentence Position task, we used the following method. First, we get sentence embeddings as described above. Then the difference between the first embedding and the other is calculated pairwise. The first embedding and its differences with others are concatenated and put as an input to a classifier.

Sentence embedding concatenation & classification: For other discourse tasks, we concatenated sentence embeddings, which were calculated as the mean of token embeddings. The concatenated sentence embeddings served as inputs for a classifier.

Masking tasks: The probing task is based on the idea of masking language modeling (see subsection 2.1). In a sentence, each word is masked, and then its probability is summed with other words' probabilities. The probability of an acceptable sentence should be higher than the probability of an unacceptable sentence. This method is for use for all tasks from BLiMP.

As a classifier model, we used logistic regression and the implementation of the Minimum Description Length method (MDL)[2] (Voita and Titov, 2020). The main reason of using MDL is the criticism that logistic regression has received lately as a probing classifier. As Voita and Titov (2020) state, logistic regression does not represent the differences in embeddings properly. For example, logistic regression might reflect the information from a corpus, not from a model itself, and be biased because of a corpus disbalance. Since our main research topic is the evaluation of language acquisition, we would have to use a method showing which part of the model's knowledge comes from the data and which part is actually acquired by it.

The main idea of MDL is to code information that comes from data and from a model separately. Since the MDL method is based on information theory, the process of learning

[2]https://github.com/alexunderch/ASR_probing

Task	Sentence examples	Labels
Subject number	*Her employer had escaped with his wife for several afternoons this summer.*	NN
	Your Mackenzie in-laws have sordid reputations few decent families wish to be connected with.	NNS
Person	*So I still can recomend them but prepare pay twice as much as they tell you initially.*	has a person marker
	The service was friendly and fast, but this just does nt make up for the lack - luster product.	does not have a person marker
Tree depth	*We have done everything we can for her .*	11
	Alvin Yeung of Civic Party	3
Top constituents	*Did it belong to the owner of the house ?*	VBD_NP_VP_.
	How long before you leave us again ?	WHNP_SQ_.
Connectors	*He 'd almost forgotten about that man . Sarah had somehow brought him back , just as she had his nightmares .*	but
	I let out a slow , careful breath . Felt tears sting my eyes .	and
Sentence position	*Quneitra Governorate (/ ALA-LC : " Muhāfazat Al-Qunaytrah ") is one of the fourteen governorates (provinces) of Syria . The governorate had a population of 87,000 at the 2010 estimate . Its area varies , according to different sources , from 685 km ² to 1,861 km ² . It is situated in southern Syria , notable for the location of the Golan Heights . The governorate borders Lebanon , Jordan and Israel .*	1
	The bossom and the part of the xhubleta covered by the apron are made out of crocheted black wool . The bell shape is accentuated in the back part . The xhubleta is an undulating , bell-shaped folk skirt , worn by Albanian women . It usually is hung on the shoulders using two straps . Part of the Albanian traditional clothing it has 13 to 17 strips and 5 pieces of felt .	4
Penn Discourse Treebank	*Solo woodwind players have to be creative,they want to work a lot*	Pragmatic Cause
	The U.S. , along with Britain and Singapore , left the agencyl, its anti-Western ideology , financial corruption and top leadership got out of hand	List
Discourse Coherence	*Within the fan inlet case , there are anti-icing air bosses and probes to sense the inlet pressure and temperature .', 'High speed center of pressure shifts along with fin aeroelasticity were major factors . At the 13th (i.e .', 'the final) compressor stage , air is bled out and used for anti-icing . The amount is controlled by the Pressure Ratio Bleed Control sense signal (PRBC) . The " diffuser case " at the aft end of the compressor houses the 13th stage .*	a text is not coherent
	This experience of digital circuitry and assembly language programming formed the basis of his book " Code : The Hidden Language of Computer Hardware and Software " . Petzold purchased a two-diskette IBM PC in 1984 for $ 5,000 . This debt encouraged him to use the PC to earn some revenue so he wrote an article about ANSI.SYS and the PROMPT command . This was submitted to PC Magazine for which they paid $ 800 . This was the beginning of Petzold 's career as a paid writer . In 1984 , PC Magazine decided to do a review of printers .	a text is coherent

Table 2: Examples of tasks

to predict labels is seen as teaching a model to transmit the data and model. Thus, the best models require less codelength to be transmitted effectively.

First, we run all experiments with logistic regression as a classifier and then we repeat the experiments with MDL. Since experiments with MDL require lots of computational resources and time, we limit them to three datasets that are better balanced and better acquired by a model based on logistic regression score. Therefore, these experiments will reveal if high results reported by logistic regression are biased because of the classifier architecture or truly represent the level of language acquisition by a transformer-based model.

4.4. Probing measures

Model	Hidden size	Number of layers	Attention heads
1	256	4	4
2	256	8	4
3	512	4	4
4	512	8	8
5	512	12	8
6	768	8	8
7	768	12	8

Table 3: Summarisation of trained models

In order to answer the question on correlation between different features' acquisition and the language levels the features belong to, we unite all the results from our experiments described above and develop standards for comparison:

- **selectivity**: selectivity is the difference in reported metrics between a pre-trained model and a random model (Hewitt and Liang, 2019). We compare checkpoints of base models in training with our baseline with shuffled labels and take the average difference (selectivity) as a measure of improvement of trained models;
- **the number of iterations needed to reach the accuracy of fully trained model**: we calculate each checkpoint's accuracy as the percentage of fully trained model's accuracy and take the first checkpoint that achieved 95%;
- **the size of a model that reaches the level of the base model**: to calculate this measure, we first train a model of the same size as the base model in the experiments before. Then we use it as a standard of comparison and train several models of different sizes

on the same data and with the same setup as the base BERT. We limit the training process to 100,000 iterations to find minimal parameters that help the model to achieve the accuracy of a base model evaluated before. Table 3 summarise models we trained to find the proper combination of parameters.

5. Results

In this section, we report results of our experiments. First, we describe outputs of experiments with two models: MultiBERT and T5, then we probe MultiBERT with the method of Minimum Length Description. The last experiment concerns the model sizes in the light of their impact on the language acquisition process.

5.1. Experiments with base models of different architectures

5.1.1. Results of MultiBERT

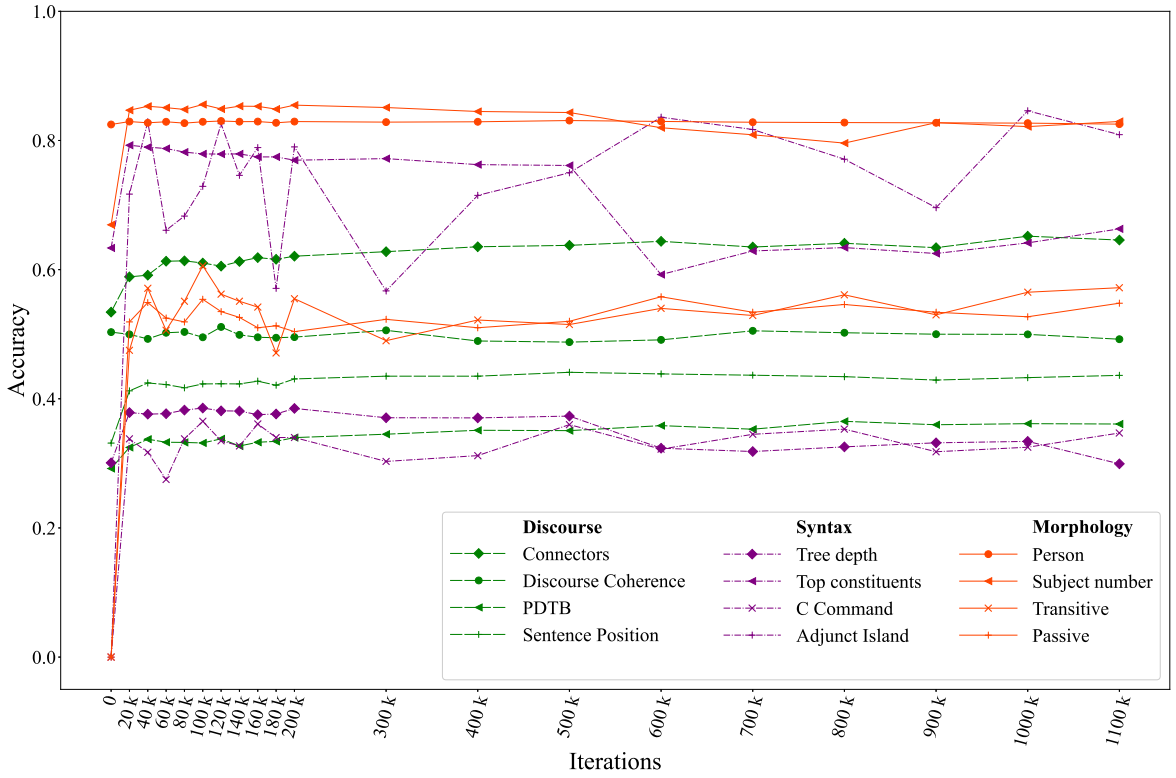

Figure 2: Comparison of MultiBERT results on different tasks.

The results of the experiments with the MultiBERT-base model are summarized in Figure 2. The model shows the best results on Subject Number and Person tasks. The classification of PDTB relations, Tree depth, and Principle A acceptability are performed with the worst accuracy.

As seen from Figure 2, accuracy of models stop changing after 600,000 iterations. However, there is a significant difference between tasks from BLiMP and other datasets. For example, the performance on the Adjunct Island task remains unstable during the whole period

of observed iterations.

Another difference between these tasks lies in the quality of models. It is illustrated with tasks grouped as ‘morphological’: Subject Number and Person tasks, which use logistic regression on MultiBERT embeddings, are solved much better than Transitive and Passive verbs. However, as follows from the plot, it is hard to group tasks based on the absolute value of accuracy.

The changing dynamics provide another perspective on model performance. From this point of view, all tasks grouped as ‘discourse’ show a similar feature: unlike others, their quality does not fluctuate but rather slightly growы across the training time. On other tasks, the model increases the quality during the first iterations. ‘Syntactic’ tasks tend to change even during later iterations. However, it is not a strict rule, and some tasks show similar behavior to ‘morphological’ ones.

5.1.2. Results of T5

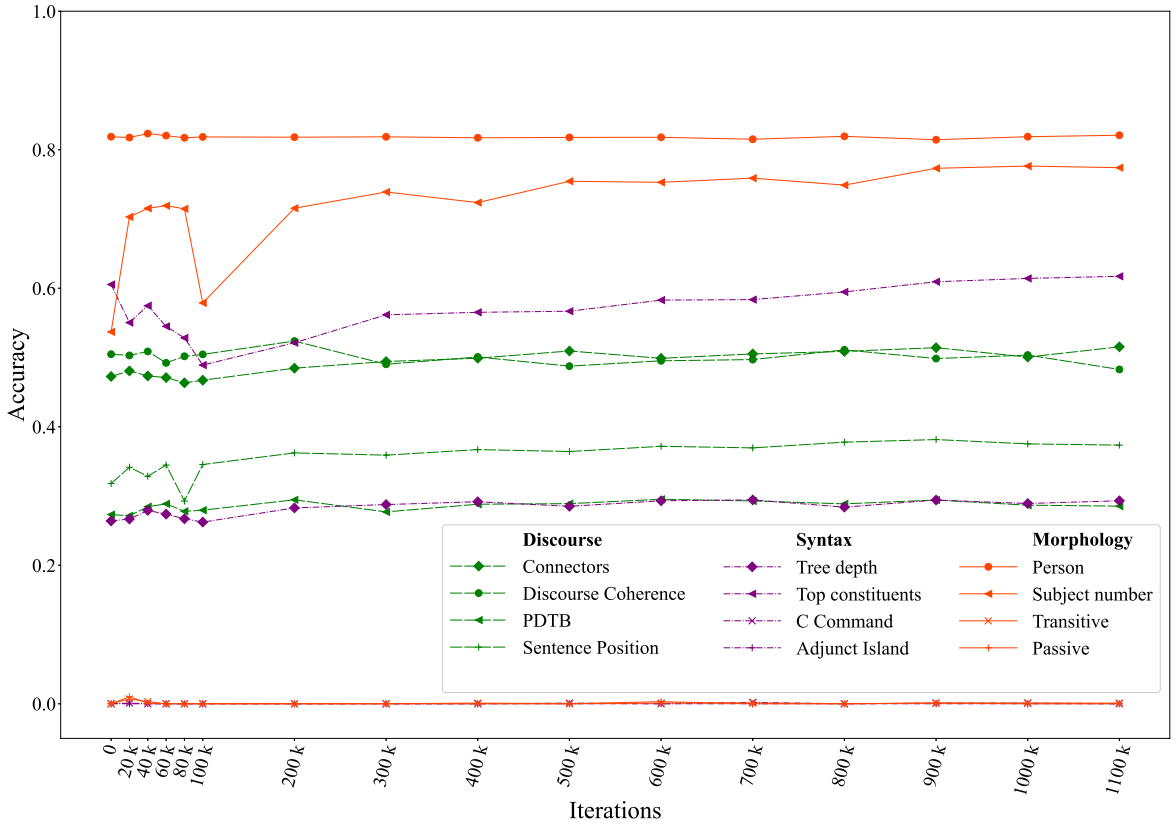

Figure 3: Comparison of T5 results on different task.

The significant difference of T5 results is the zero-close quality on BLiMP datasets due to the architecture and training process designed for generation. Except for these tasks, the quality of T5 is similar to MultiBERT. The best performance is on the Person task, and the worst quality is shown on PDTB relation classification and Tree depth.

Unlike MultiBERT, we first used the available checkpoints of T5 with a step of 100,000 iterations. Then we trained a new model on the same resources and texts, but it might be a better initialization, which affected the final results.

Similar to MultiBERT, 'discourse' tasks show almost no significant change and slow growth, whereas the models' performance increases on 'syntactic' and 'morphological' tasks during the first 100,000 iterations.

5.1.3. Comparison of models of different architectures

We described the surface results of models' performance and now can dive deep into more detailed results. The results described above should be considered relative. To illustrate how much information models acquire during these iterations, we compare them to final models. As the process of training T5 was not finished, we compare this model with the original T5 (Raffel et al., 2019). As seen from Figure 4, MultiBERT scores are close to results of the final checkpoint. Hence, there is no need to look at later iterations. The comparison with the original T5 shows that the model we use performs worse due to the smaller resources it was trained on. Therefore, the difference in quality should not be explained by the difference in architecture.

However, we should consider that some tasks are solved with the same quality as a baseline with shuffled labels (Discourse Coherence and Person). Moreover, T5 does not perform much better than the baseline on Penn Discourse Treebank relations. Consequently, models encounter difficulty with 'discourse' tasks.

Furthermore, MultiBERT and T5 show similar learning trajectories on several tasks, such as Connectors and Sentence Position tasks. Another key feature shared by the two models is the full termination of increases between 500,000 and 600,000 iterations. Despite the fact that models vary in size and training process, they show some similarities in probing tasks. Hence, the acquisition generally does not depend on the model architecture.

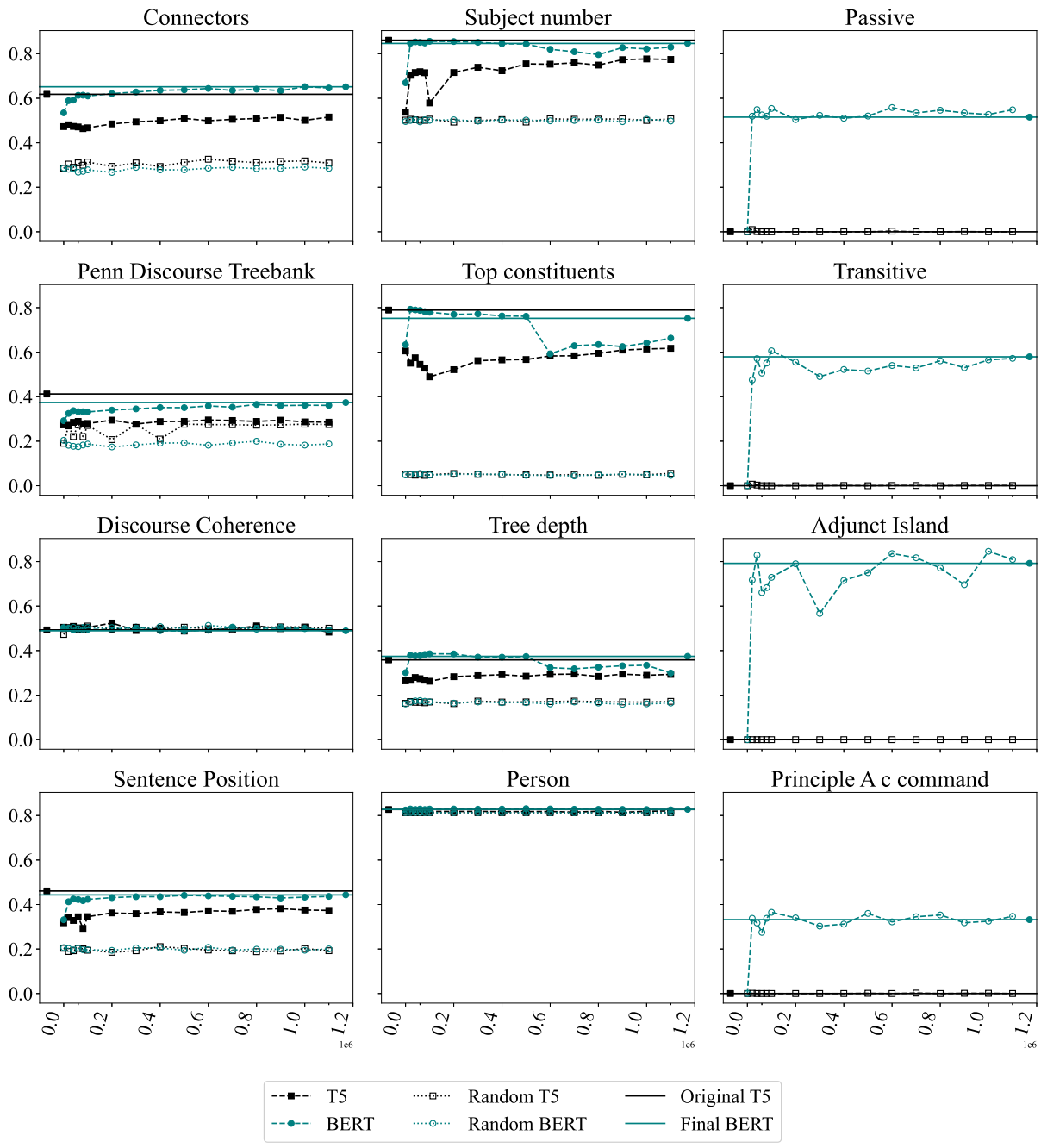

Figure 4: Performance on models on different tasks.

5.2. *Experiments with Minimum Length Description*

As previous experiments revealed no difference in architectures, we limit these and forthcoming experiments to BERT-based models.

After running experiments with logistic regression as a diagnostic classifier, we choose

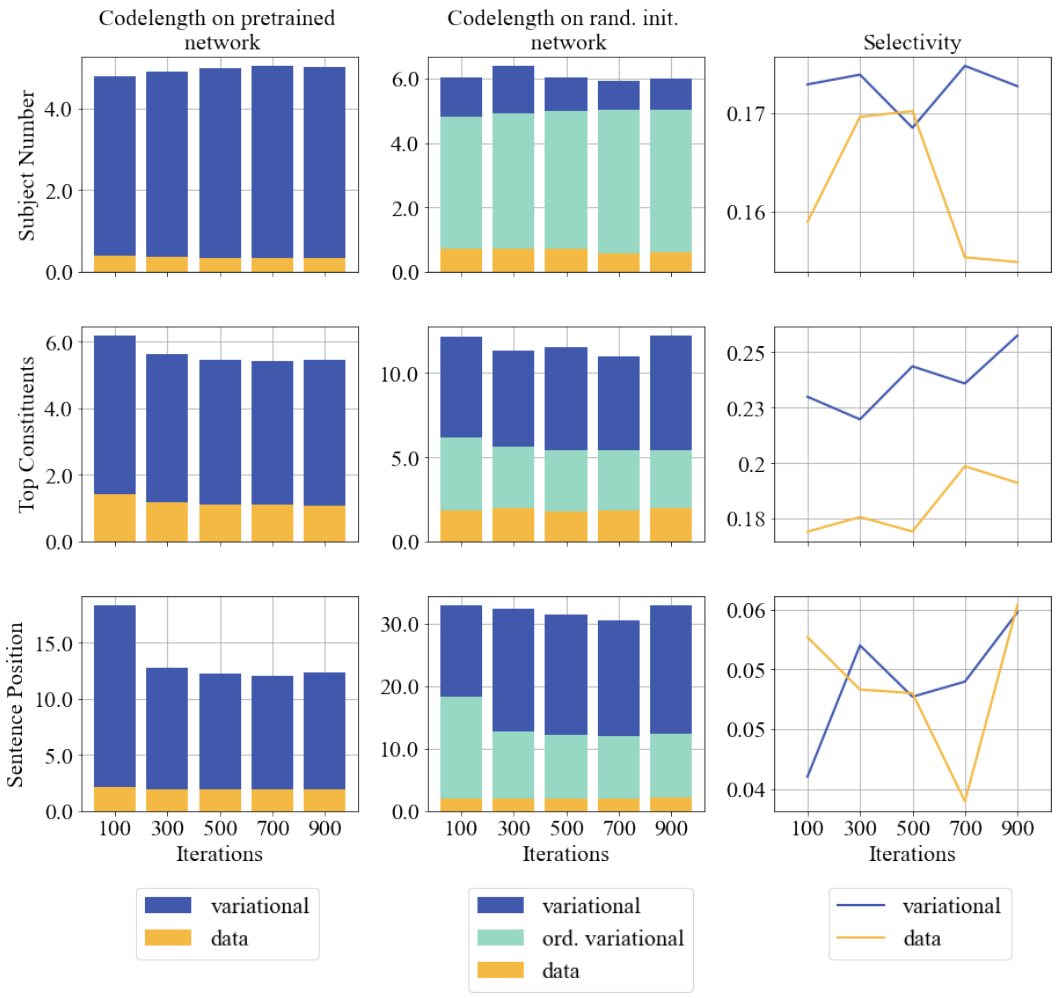

Figure 5: Minimum Description Length on three tasks.

three tasks, one from each language level, with the highest score and the best dataset balance (based on Figure 7). Therefore, we run experiments with MDL on Sentence Position, Subject Number and Top Constituents tasks. The results are summarized in Figure 5. The first two columns describe codelength of pre-trained and randomly initialised models respectively, and in the third column selectivity of metrics (weighted F1-score for these experiments) is reported.

As can be seen from the chart, Subject Number shows the best quality, since it has the shortest codelength. However, a randomly initialised model shows good results too, which means that this task is simple enough that a randomly initialised model can solve it with relatively good quality. Moreover, if we look at the performance by layers (see Figure 8 in Appendix), we conclude that the difference in layers' metrics is not smoothing, as we could

expect.

On Top Constituents task a model has reached small codelength by the last epoch, and the selectivity is even higher than for Subject Number. Regarding codelength, a pre-trained model also shows better results than a randomly initialised model. The quality of all layers has increased by the 900,000 iterations, as Figure 9 shows.

Unlike previous tasks, Sentence Position task seems not to be acquired. First, a model has very low selectivity. Second, the codelength is big even on the last iterations in comparison with other tasks. However, the quality is increasing significantly across the iterations. That could mean that this task is too difficult for this architecture. Unlike the model's behaviour on other tasks, the model has room for improvement, and the model is trying to adjust to solve the task better. As for layers' quality across iterations, it does not change significantly.

Overall, MDL shows similar results to logistic regression. MDL also reveals that little to no difference can be seen in codelength results of iterations after 500,000 training steps. Subject Number and Top Constituents tasks show a little decrease in codelength even between 200,000 and 400,000 iterations.

5.3. Experiments with models' sizes

The previous results show that models of different architectures behave similarly and the acquisition of most of the linguistic features stops after 500,000 steps. Therefore, to find out which model parameters affect the score, our next step is to train small models to look closer at the first training steps. First, we conducted the same experiments on four small models described in section 4.1.

Compared to MultiBERT, models show worse accuracy. However, among small models, the one with the increased hidden size shows the best results in all cases, except for Penn Discourse Treebank and Tree depth, where the model with the increased number of layers shows the best results. This model shows the second best results on other tasks.

The behaviour of the model with the increased number of attention heads is inconsistent compared to the tiny model (hidden size of 128, 2 attention heads, and 2 layers). On some tasks, such as Penn Discourse Treebank and Discourse Coherence, it shows worse accuracy than the tiny model. On other tasks, it shows better quality than the tiny model but worse than other models.

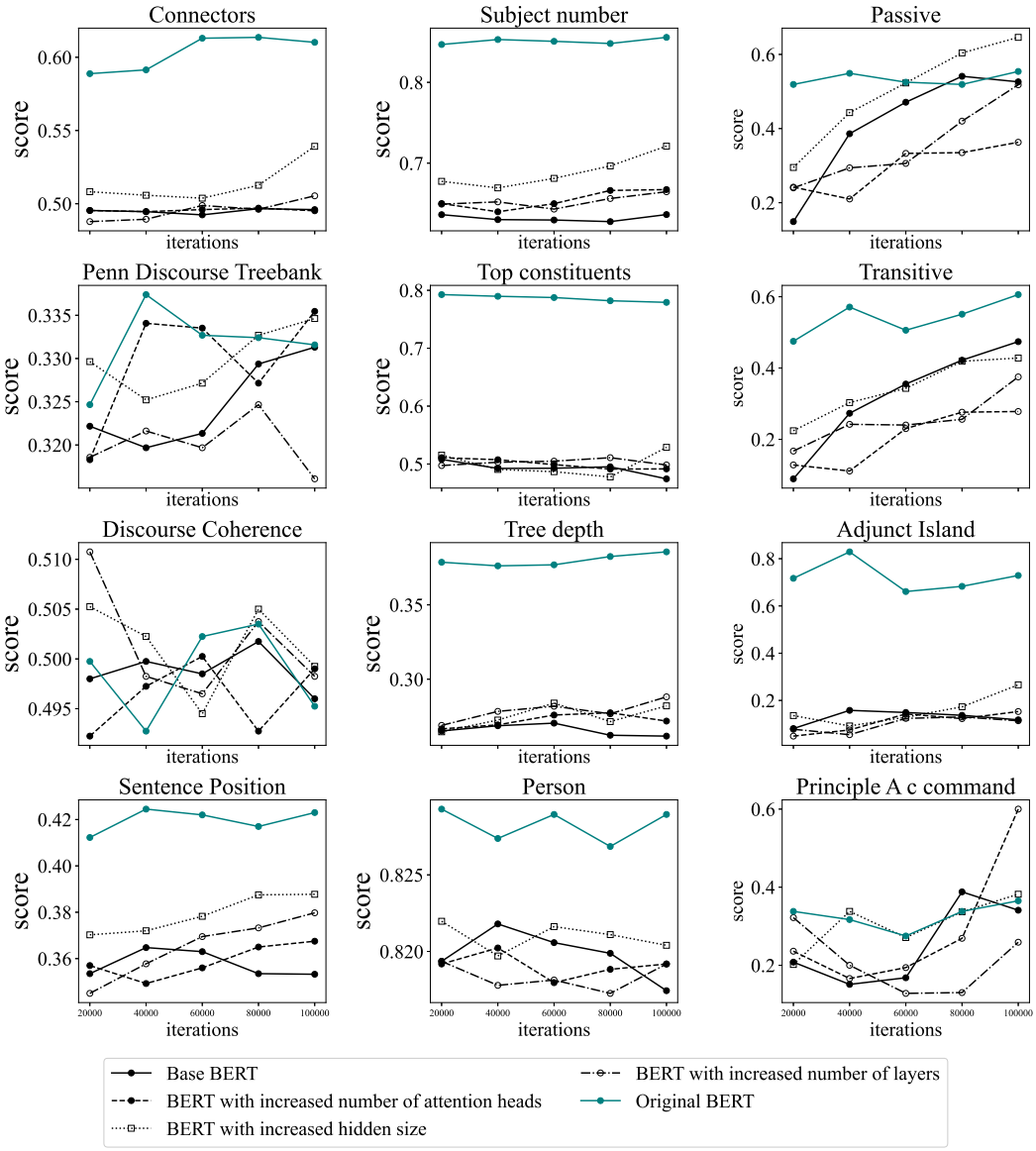

Figure 6: Small models' results on different tasks.

Nevertheless, these observations are not applicable to the tasks from BLiMP. As charts show, on tasks, such as Passive and Principle A, the tiny model shows better quality than any other models, including the MultiBERT model. At the same time we see that small models encounter difficulties with the acceptability of sentences with transitive verbs and with islands.

The described above leads to the conclusion that bigger models are more successful in language acquisition. It is worth mentioning that different parameters of model size give different level of improvement. Thus, the most important parameter for language acquisition

Level	Task	BERT, avg Δ	T5, avg Δ	Iterations needed	Model size
morphology	subject number	0.304	0.25	20000	768/8
morphology	person	0.017	0.01	0	128/2
syntax	top constituents	0.636	0.576	20000	768/8
syntax	tree depth	0.181	0.129	20000	768/8
discourse	discourse coherence	-0.002	0.0	0	128/2
discourse	Connectors	0.321	0.237	200000	768/8
discourse	Sentence Position	0.208	0.164	40000	512/4
discourse	Penn Treebank	0.149	0.074	600000	128/4

Table 4: The comparison of tasks' acquisition

is hidden size, since it leads to better results for most features. The second best parameter is the number of layers.

5.4. Correlation between levels of morphology and tasks

We measured the quality of models with different criteria described in section 4.4. As seen from Table 4, discourse takes many more iterations to reach the level of the fully trained model (we compared the results on BERT, as the T5 model we used for our experiments was not trained until later epochs). 'Morphological' and 'syntactic' features reach the 95%-level of the fully trained model much faster than 'discourse' tasks.

The average selectivity reported in columns 3 and 4, i. e. difference between a random baseline and trained models, do not show any difference between different language levels, except for two tasks discussed earlier - Discourse Coherence and Person, which were not acquired by models at all, since trained models show the same accuracy as the randomly initialised model.

The last criterium concerns the size of a model that shows similar quality to the base model. Following the results we achieved in our experiments with model parameters, we limited our experiments to hidden size and number of layers leaving behind the number of attention heads as an insignificant factor. These experiments show that most of the 'morphosyntactic' tasks are acquired by models with hidden size of 768. At the same time, on discourse-based tasks, models with much smaller size show results comparable to the base model.

As mentioned above, BLiMP tasks were solved in an unsupervised manner, therefore, we cannot compare them with our baseline, as it is based on shuffled labels. In table 5, we

Level	Task	BERT, avg	T5, avg	Iterations needed	Model size
morphology	passive	0.503	0.001	20000	512/4
morphology	transitive	0.514	0.001	40000	512/8
syntax	adjunct island	0.711	0.0	40000	768/8
syntax	principle A	0.514	0.001	20000	512/4

Table 5: The comparison of BLiMP tasks' acquisition

summarise model performance on these tasks. Instead of difference between accuracy and our baseline, we use average accuracy. Similar to other tasks on morphology and syntax, models reach the 95%-level of the fully trained model's performance around 20,000-40,000 iterations. However, the experiments with models of different sizes show slightly different results. For most BLiMP tasks the level of the base models is achieved by models of hidden size of 518 with 4 or 8 layers, which is a smaller size than for other 'morphosyntactic' probing tasks.

Regarding accuracy on all tasks, it does not show any difference in tasks that belong to different levels of language structure.

6. Discussion

In this section, we use the results of our experiments to answer the questions posed in the introduction: (i) do models acquire language and if they do, how many training steps does it take? (ii) which parameters are crucial for language acquisition? (iii) is there any correlation between language levels and the acquisition process?

6.1. Do models acquire language and how fast?

The experiments with both logistic regression and MDL show that linguistic information is acquired fast, before 600,000 training iterations. It corresponds to results of other researchers (Blevins et al., 2022; Chiang et al., 2020; L. Z. Liu et al., 2021) that independently showed similar results on a fast acquisition of linguistic features. Discourse is not fully acquired by the end of the observed training period, compared to the baseline results. The difference between 'syntactic' and 'morphological' tasks is insignificant, which correlates with ideas on morphosyntax (Caha, 2009; Embick and Noyer, 2007). Despite the fact that we cannot prove that models regard morphology and syntax as the same level, we can make a less strict statement that models acquire these grammatical units simultaneously.

BLiMP gives another perspective on the process of acquisition. MultiBERT results remain unstable for a longer period than similar tasks with classifiers. It might indicate the difference between two different approaches to probing. However, from the linguistic point of view, BLiMP includes more difficult linguistic features cases, while SentEval tasks evaluate more basic knowledge. Besides that, T5 architecture does not allow to use this dataset in the same way as for MultiBERT since Masking Language Modeling and T5 generation are different tasks. Therefore, we cannot compare language acquisition of two models on these tasks.

Apart from that, we can state that the model architecture does not effect the quality of language acquisition, and both models show basic linguistic knowledge.

6.2. What parameters of models help to acquire language?

The results of our experiments with model sizes show that the increase of hidden size has the biggest impact on the quality of models. The number of layers was the second important parameter and improved quality better than the number of attention heads. Our results are similar to the results reported in Z. Wang et al. (2019): they also showed that larger hidden size tended to improve quality.

The hidden size might be important for smaller models because different layers code different information. For example, Rogers et al. (2020) summarised that the first layers are task-invariant and contain general linguistic information while the latest layers are usually task-specific.

On the contrary, attention heads are usually more detailed, for example, they are known to remember specific syntactic patterns (Htut et al., 2019). Kovaleva et al. (2019) revealed that attention heads learn the same patterns. Therefore, when the resources to encode information are limited, attention heads do not add much new information.

Regarding the hidden size, our results are different from the results in (Z. Wang et al., 2019). While they postulate that number of layers is the most essential parameter, our results show that hidden size is better for performance improvement.

The results of the experiments reported in tables 5 and 4 prove that increasing hidden size shows better results than increasing number of layers. Moreover, models with the hidden size of 768 and 8 layers show results close to the model with the same hidden size and 12 layers. Therefore, we conclude that hidden size is the crucial parameter for language acquisition.

6.3. Correlation between language levels and model's acquisition

Our results reveal that there is not much difference in acquisition of morphological and syntactic features. However, most of the experiments show that discourse-based tasks are much slower to acquire, and models show lower accuracy on these tasks. Therefore, we can conclude that models are able to distinguish language levels.

Whether models acquire discourse and to what extent remain open questions. The MDL experiment on 'discourse' task shows very low selectivity and a significant difference in code-length of a pre-trained model and a randomly initialised model (see Figure 5). It shows that while the loss of the model is decreasing, which is shown by codelength, it does not lead to the better performance.It might indicate that BERT-based model is not capable of learning discourse on a high level.

Beside that, models of smaller sizes acquire discourse features on the same level as the base model (see Table 4). That might be interpreted in two ways. On the one hand, it could mean that discourse does not require so much 'encoding ability', and it requires other kind of knowledge being acquired first before models would show better results on discourse. On

the other hand, it might highlight that the change of architecture parameters does not help in discourse acquisition. We leave testing these hypotheses for future research.

There are a few possible explanations for worse performance on discourse than on other language levels. First, while working with discourse, logistic regression gets longer embeddings than for other tasks (see subsection 4.3 on methods), which might effect the model performance because it has to process more information. However, the correlation between embedding size and the level of performance requires additional research.

The second explanation concerns the nature of discourse. The discourse itself is closer to pragmatics and extralinguistic knowledge. Moreover, discourse requires the understanding of semantics, which is learnt throughout the training process (L. Z. Liu et al., 2021). On the opposite, morphology and syntax include low-level structures, so models can acquire them faster.

On the contrary, morphology and syntax are deeply connected. To detect a subject number, which was the task attributed to morphology, a model has to restore a syntactic hierarchy to find a subject. To find top of constituents in the syntactic tree, a model has to classify them by parts of speech. This classification is based not only on syntactic features but on morphological ones as well. Despite the fact that BLiMP tasks show similar results on morphology and syntax (see Table 5), the learning curve of syntactic features is fluctuating during the training process, as can be seen from Figures 2 and 3.

7. Conclusion

This thesis addresses the problem of language acquisition in state-of-the-art models and answers several research questions: when models acquire the linguistic information, how different grammar categories are acquired in comparison with one another and which factors influence the language acquisition process.

As our results show, the linguistic information is acquired pretty early. While the process of model training includes more than 1,000,000 steps, morphology and syntax seem to be learnt during the first 600,000 steps. The discourse takes longer time to be acquired. Whereas morphology and syntax tasks show the similar acquisition patterns, discourse is significantly different, as results on discourse tend to be lower and might not be learnt at all.

Regarding architectures of models, both T5 and MultiBERT demonstrate comparable results considering the quality of the language level acquisition. T5 does not yield any results on BLiMP due to the generation algorithm. Most tasks show that T5 acquires basic morphological and syntactic features and some discourse features. MultiBERT does not improve its quality on some discourse tasks compared to randomly labeled embeddings. However, it could be said that MultiBERT acquires each level to some extent.

We also experimented with the Minimum Length Description method to check whether it would show any difference with results of logistic regression. We find out that the 'discourse'-based task is barely solved, while the model shows good quality on 'morphological' and 'syntactic' tasks.

To display correlation between language acquisition and different model parameters, we trained four models: one with the minimal hidden size and minimal number of layers and attention heads and three models with one parameter increased and others frozen. These experiments reveal that hidden size appears to be the most essential parameter for language acquisition, whereas attention heads do not significantly increase a model's performance.

Finally, we compared all tasks with several criteria: selectivity (the difference between pre-trained and randomly initialised models), the number of iterations needed to reach the level of fully trained model, and the size of a model that shows the quality comparable with the base model used before. The idea behind this comparison is to find any correlation between different language levels and probing measures. As a result, models distinguish discourse from morphology and syntax but there is almost no difference between 'morphological' and

'syntactic' tasks.

As a practical result of this research, we present a new publicly available framework for language acquisition probing methodology[3].

[3]https://github.com/EkaterinaVoloshina/chronological_probing

References

Belinkov, Y., Gehrmann, S., & Pavlick, E. (2020). Interpretability and analysis in neural nlp. *Proceedings of the 58th annual meeting of the association for computational linguistics: tutorial abstracts*, 1–5.

Blevins, T., Gonen, H., & Zettlemoyer, L. (2022). Analyzing the mono- and cross-lingual pretraining dynamics of multilingual language models. https://doi.org/10.48550/ARXIV.2205.11758

Caha, P. (2009). The nanosyntax of case.

Chen, M., Chu, Z., & Gimpel, K. (2019). Evaluation benchmarks and learning criteria for discourse-aware sentence representations. *Proc. of EMNLP*.

Chiang, C.-H., Huang, S.-F., & Lee, H.-y. (2020). Pretrained language model embryology: The birth of albert. *arXiv preprint arXiv:2010.02480*.

Chomsky, N. (1965). Aspects of the theory of syntax cambridge. *Multilingual Matters: MIT Press*.

Chomsky, N. (1981). Lectures on government and binding (dordrecht: Foris). *Studies in generative grammar*, *9*.

Choshen, L., Hacohen, G., Weinshall, D., & Abend, O. (2022). The grammar-learning trajectories of neural language models. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8281–8297. https://aclanthology.org/2022.acl-long.568

Clark, A., & Lappin, S. (2010). Computational learning theory and language acquisition. *Philosophy of linguistics*, 445–475.

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., & Baroni, M. (2018). What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*.

Crystal, D. (1987). Towards a 'bucket'theory of language disability: Taking account of interaction between linguistic levels. *Clinical Linguistics & Phonetics*, *1*(1), 7–22.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Embick, D., & Noyer, R. (2007). Distributed morphology and the syntax/morphology interface. *The Oxford handbook of linguistic interfaces*, *289324*.

Hewitt, J., & Liang, P. (2019). Designing and interpreting probes with control tasks. *arXiv preprint arXiv:1909.03368*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Htut, P. M., Phang, J., Bordia, S., & Bowman, S. R. (2019). Do attention heads in bert track syntactic dependencies? *arXiv preprint arXiv:1911.12246*.

Kovaleva, O., Romanov, A., Rogers, A., & Rumshisky, A. (2019). Revealing the dark secrets of bert. *arXiv preprint arXiv:1908.08593*.

Krause, B., Kahembwe, E., Murray, I., & Renals, S. (2019). Dynamic evaluation of transformer language models. *arXiv preprint arXiv:1904.08378*.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Lewis, J. D., & Elman, J. L. (2001). Learnability and the statistical structure of language: Poverty of stimulus arguments revisited. *Proceedings of the 26th annual Boston University conference on language development*, *1*, 359–370.

Liu, L. Z., Wang, Y., Kasai, J., Hajishirzi, H., & Smith, N. A. (2021). Probing across time: What does roberta know and when? *arXiv preprint arXiv:2104.07885*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Marcus, M., Kim, G., Marcinkiewicz, M. A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., & Schasberger, B. (1994). The penn treebank: Annotating predicate argument structure. *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

McCoy, R. T., Min, J., & Linzen, T. (2019). Berts of a feather do not generalize together: Large variability in generalization across models with similar test set performance. *arXiv preprint arXiv:1911.02969*.

Mikhailov, V., Serikov, O., & Artemova, E. (2021). Morph call: Probing morphosyntactic content of multilingual transformers. *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, 97–121. https://doi.org/10.18653/v1/2021.sigtyp-1.10

Nie, A., Bennett, E., & Goodman, N. (2019). Dissent: Learning sentence representations from explicit discourse relations. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4497–4510.

Niu, T., Yavuz, S., Zhou, Y., Wang, H., Keskar, N. S., & Xiong, C. (2020). Unsupervised paraphrase generation via dynamic blocking. *arXiv preprint arXiv:2010.12885*.

Prefors, A., Regier, T., & Tenenbaum, J. B. (2006). Poverty of the stimulus? a rational approach. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *28*(28).

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, *8*, 842–866.

Ross, J. R. (1967). Constraints on variables in syntax.

Saphra, N. (2021). Training dynamics of neural language models.

Saphra, N., & Lopez, A. (2018). Understanding learning dynamics of language models with svcca. *arXiv preprint arXiv:1811.00225*.

Sellam, T., Yadlowsky, S., Wei, J., Saphra, N., D'Amour, A., Linzen, T., Bastings, J., Turc, I., Eisenstein, J., Das, D. et al. (2021). The multiberts: Bert reproductions for robustness analysis. *arXiv preprint arXiv:2106.16163*.

Van Bezooijen, R., & Gooskens, C. (1999). Identification of language varieties: The contribution of different linguistic levels. *Journal of language and social psychology*, *18*(1), 31–48.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Voita, E., & Titov, I. (2020). Information-theoretic probing with minimum description length. *arXiv preprint arXiv:2003.12298*.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Wang, Z., Mayhew, S., Roth, D. et al. (2019). Cross-lingual ability of multilingual bert: An empirical study. *arXiv preprint arXiv:1912.07840*.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. R. (2020). Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, *8*, 377–392.

Zhang, Y., Warstadt, A., Li, H.-S., & Bowman, S. R. (2020). When do you need billions of words of pretraining data? *arXiv preprint arXiv:2011.04946*.

Appendix A. The dataset balance

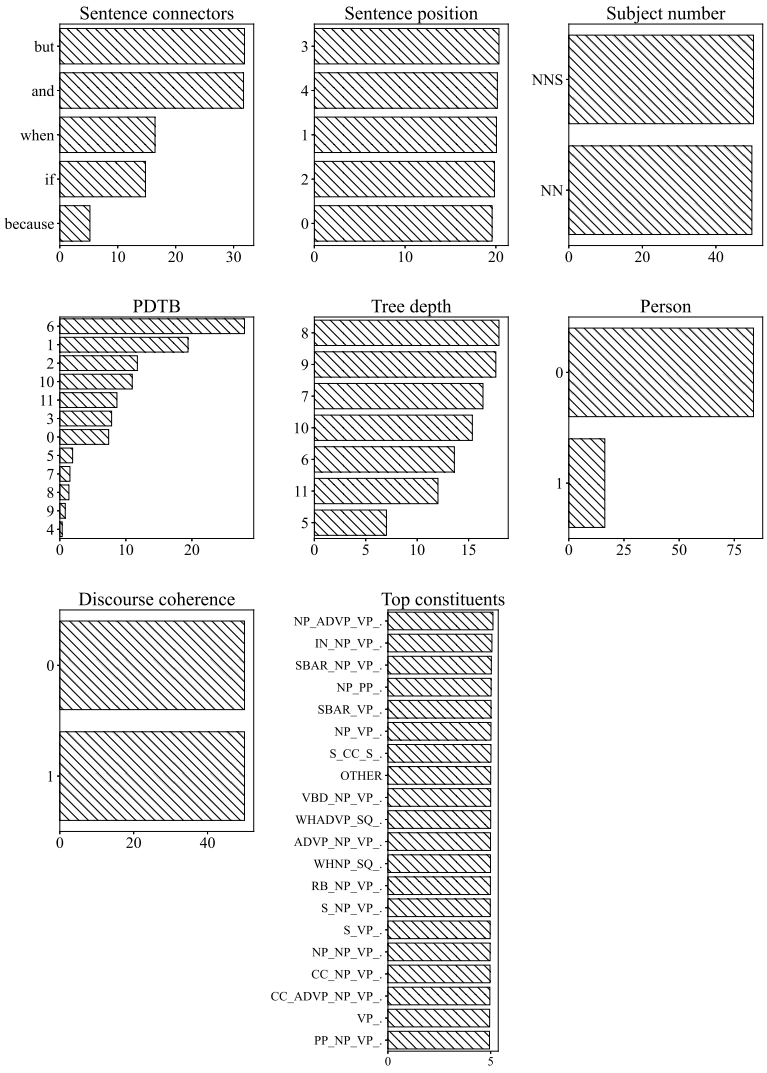

Figure 7: The balance of datasets.

Appendix B. The results of experiments with Minimum Description Length

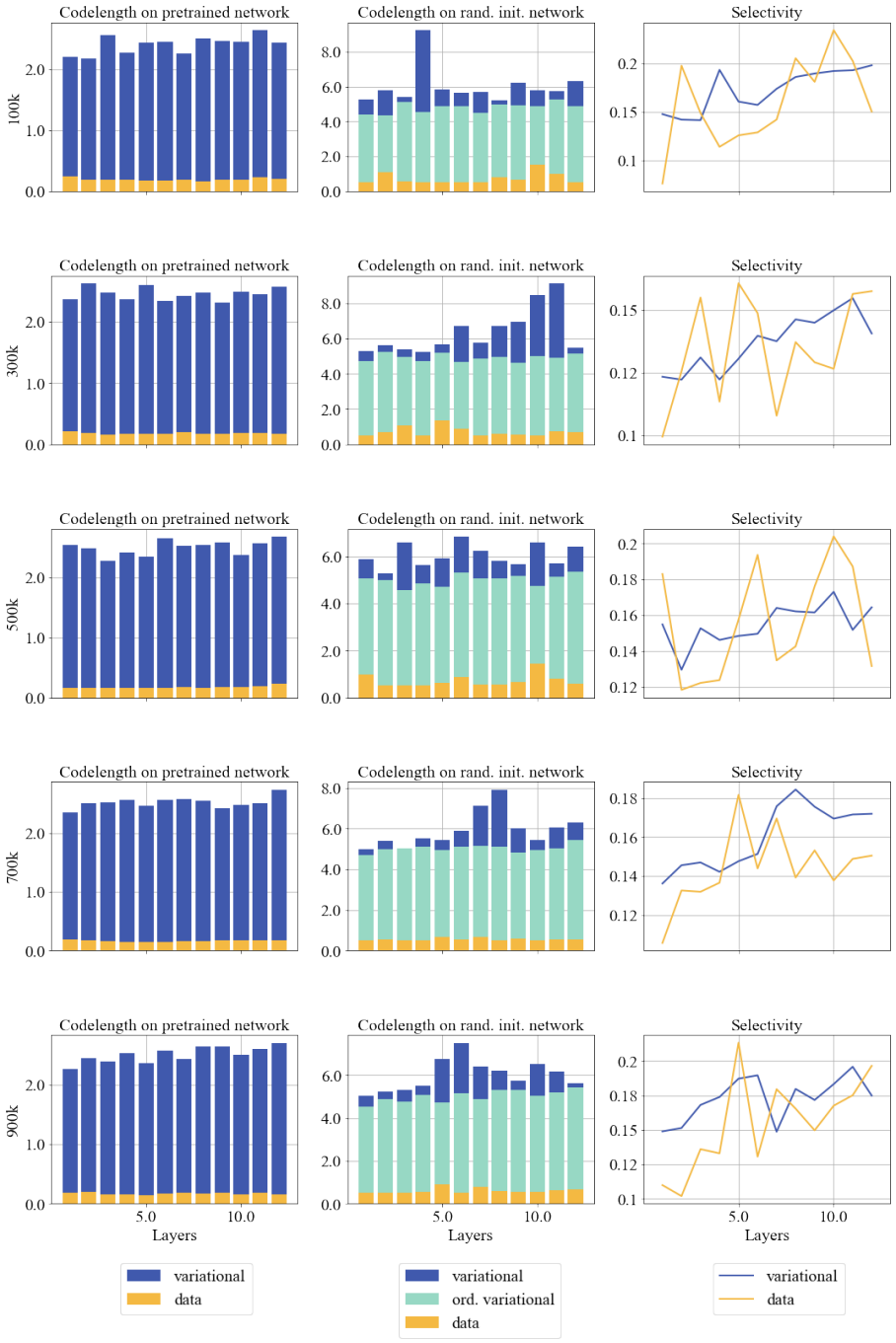

Figure 8: Minimum Description Length on Subject Number task.

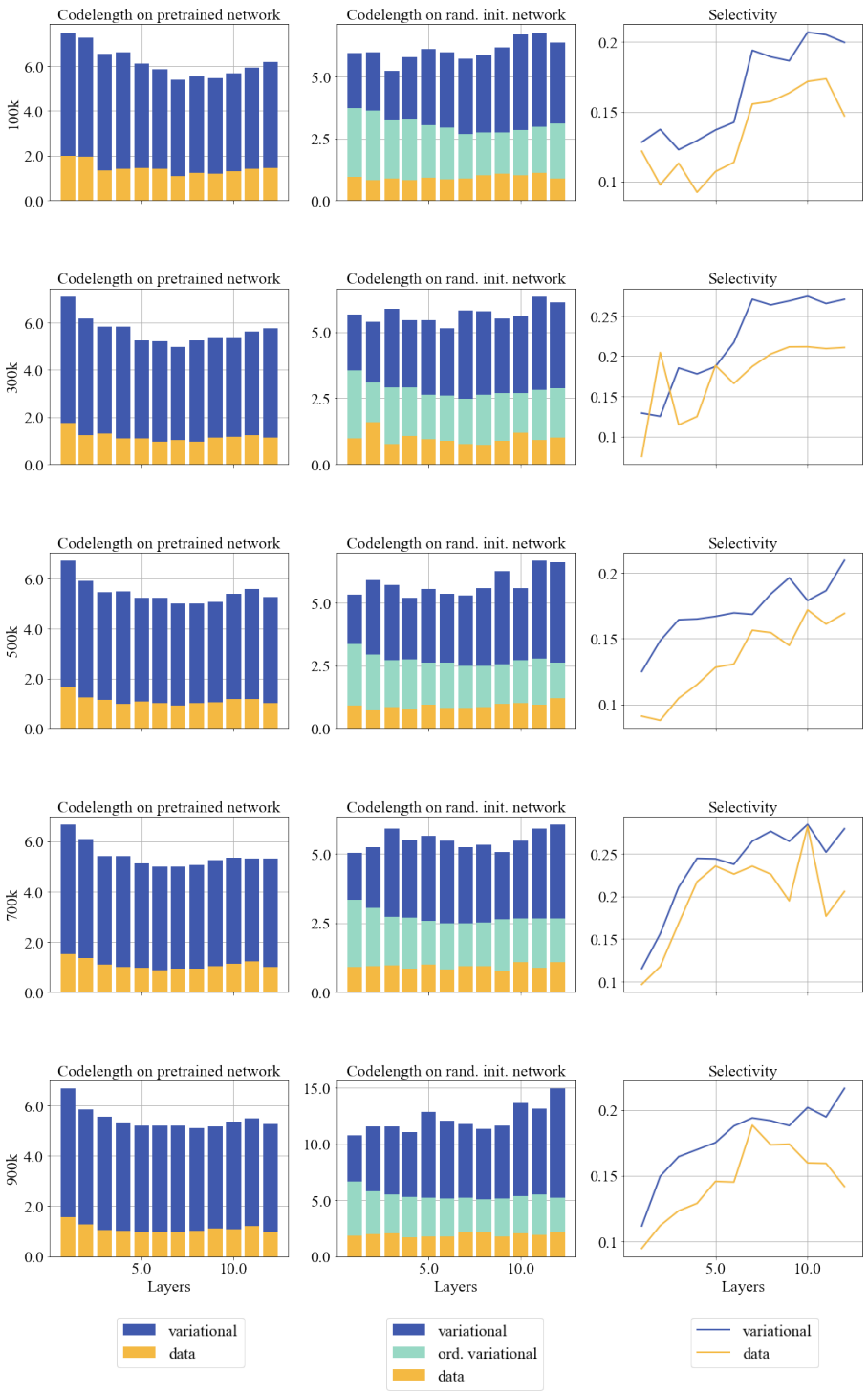

Figure 9: Minimum Description Length on Top Constituents task.

Figure 10: Minimum Description Length on Sentence Position task.