On the Nature of BERT: Correlating Fine-Tuning and Linguistic Competence

Federica Merendi[⋆], **Felice Dell'Orletta**[◇], **Giulia Venturi**[◇]
[⋆]Department of Computer Science, University of Pisa
[◇]Istituto di Linguistica Computazionale "Antonio Zampolli", Pisa
ItaliaNLP Lab – *www.italianlp.it*
f.merendi@studenti.unipi.it, {name.surname}@ilc.cnr.it

Abstract

Several studies in the literature on the interpretation of Neural Language Models (NLM) focus on the linguistic generalization abilities of pre-trained models. However, little attention is paid to how the linguistic knowledge of the models changes during the fine-tuning steps. In this paper, we contribute to this line of research by showing to what extent a wide range of linguistic phenomena are forgotten across 50 epochs of fine-tuning, and how the preserved linguistic knowledge is correlated with the resolution of the fine-tuning task. To this end, we considered a quite understudied task where linguistic information plays the main role, i.e. the prediction of the evolution of written language competence of native language learners. In addition, we investigate whether it is possible to predict the fine-tuned NLM accuracy across the 50 epochs solely relying on the assessed linguistic competence. Our results are encouraging and show a high relationship between the model's linguistic competence and its ability to solve a linguistically-based downstream task.

1 Introduction

In the last few years, interest in assessing the linguistic generalization abilities of Neural Language Models (NLMs) has given rise to numerous studies aimed at investigating how the models are able to encode different types of linguistic phenomena. To this end, the most widespread methodology is the *probing classification approach* (Conneau et al., 2018; Warstadt et al., 2019; Hewitt and Liang, 2019), which showed that NLMs are able to encode a variety of language properties (Rogers et al., 2020). Even if most of the work is focused on the linguistic competence of pre-trained models, a complementary line of research is devoted to understanding whether and to what extent the existing competence is modified by fine-tuning the pre-trained model on a supervised downstream dataset (Merchant et al., 2020).

In this paper, we continue this line of research and carry out a study aimed at adopting a probing task approach to investigate how the linguistic competence of one of the most prominent NLM model, BERT (Devlin et al., 2019), is altered after a fine-tuning stage with the main focus on which specific phenomena are preserved or forgotten and which fine-tuning epochs are mainly involved. As a testbed, we chose a downstream task where linguistic information plays the main role. It consists in predicting the evolution of written language competence relying on the linguistic style of chronologically ordered productions written by language learners (Richter et al., 2015; Miaschi et al., 2021a). To the best of our knowledge, it has been never considered in the NLM interpretability literature, and differently from previous studies, the task was carried out at sentence rather than at document level. The present study addresses two further related issues: the first one concerns the debated relationship between the linguistic competence of an NLM and its ability to resolve a task (Ravichander et al., 2021). Secondly, we aim to investigate whether it is possible to predict the accuracy of a fine-tuned model across multiple training epochs solely relying on the linguistic probing task results reflecting BERT's linguistic competence.

Contributions. In this paper, we *i)* compared the ability of BERT's and a Support Vector Machine classifier, which uses a set of explicit linguistic features, in the prediction of the chronological order of two sentences written by the same student across two school years, *ii)* showed the impact of the fine-tuning stage on BERT's linguistic competences, focusing in particular on how they change across 50 epochs of fine-tuning, *iii)* assessed which types of linguistic competence are mainly related to BERT's ability to solve the downstream task, *iv)* predicted the fine-tuned BERT's performance across 50 epochs relying on the linguistic competence of the model.

Proceedings of the 29th International Conference on Computational Linguistics, pages 3109–3119
October 12–17, 2022.

2 Related work

Several methods have been proposed in the literature to obtain meaningful explanations of how NLMs are able to capture syntax- and semantic-sensitive phenomena (Belinkov et al., 2017), also taking inspiration from human language experiments (Ettinger, 2020). Despite highly debated (Belinkov, 2021), one of the mostly explored methods is the definition of *probing tasks* which a model can solve only if it has encoded a precise linguistic phenomenon within its representations.

Particularly relevant to our study are the multiple papers focused on the impact of the fine-tuning stage on how and the extent to which the linguistic knowledge encoded in the pre-trained model is modified, such as Mosbach et al. (2020); Merchant et al. (2020); Durrani et al. (2021); Sarti et al. (2021); Miaschi et al. (2020); Yu and Ettinger (2021) to mention the most recent ones. Even with some main differences concerning the NLMs, and the fine-tuning and probing tasks considered, they agree that the main changes are typically larger for the output layers of a fine-tuned model than for the first ones. However, a shared consensus about whether the linguistic information are preserved, reinforced or forgotten is still missing. For example, Mosbach et al. (2020); Merchant et al. (2020); Sarti et al. (2021) report both a general not catastrophic forgetting and in some cases a substantial improvement of the linguistic competence directly involved in the resolution of the downstream task. On the contrary, Yu and Ettinger (2021) found that fine-tuning on phrase meaning composition sets does not exhibit noteworthy benefits, and Miaschi et al. (2020) show that BERT tends to lose its precision in encoding a wide set of linguistic features after the fine-tuning process. However, to the best of our knowledge, none of the prior works provide a comprehensive analysis of the impact of the fine-tuning stage across the training epochs.

An orthogonal debated issue we are going to address in this study regards the question raised by Ravichander et al. (2021) and concerning the extent information encoded in NLMs is indicative of information needed to perform downstream tasks. Differently from the authors who demonstrated that NLMs are able to encode linguistic properties even if they are not needed for a given downstream task, our results show that there is a strong correlation between the ability to encode a linguistic property and the accuracy to solve the task. An additional novelty of this study is the language we focus on, i.e. Italian. While the vast majority of these studies are focused on English, relatively little work has been done to interpret non-English models. Exceptions are represented by de Vries et al. (2020) focused on Dutch and by Miaschi et al. (2021b); Guarasci et al. (2022) dealing with Italian.

3 Our Approach

Our study includes multiple experiments. Firstly, we test BERT's ability to solve the downstream task, by comparing the performance of the fine-tuned model against a Support Vector Machine classifier which uses a set of explicit features resulted particularly relevant in the resolution of the task, as shown by Richter et al. (2015); Miaschi et al. (2021a). Then, we use a suite of probing tasks to test the linguistic abilities of pre-trained and fine-tuned BERT, with a specific focus on how they change across multiple training epochs of fine-tuning and on which linguistic properties are mostly correlated with the resolution of the downstream task. As linguistic probing tasks, we use the same set of linguistic features used by the SVM. The last experiment is devoted to assess whether it is possible to predict the fine-tuning performance using the accuracy of the linguistic probing tasks.

3.1 Data

We used two datasets: (i) the Universal Dependency Italian Treebank (IUDT) for probing the linguistic knowledge learned before and after a fine-tuning process; (ii) a collection of chronologically ordered essays contained in the *CItA* corpus, which we used to solve the predicting evolution task and for fine-tuning.

IUDT dataset. It includes all 5 sections of the Italian treebank built in the framework of the UD project (de Marneffe et al., 2021), version 2.5, for a total of 33,017 sentences.

CItA (Corpus Italiano di Apprendenti L1). The corpus is the first longitudinal collection of productions written by first language (L1) learners existing for the Italian language (Barbagli et al., 2016). It includes 1,352 essays written by a total of 156 students, aged 11-12, followed from the first to the second year of four different Italian lower secondary schools. Note that this temporal span is considered quite crucial for Italian L1 learners, since a number of relevant transforma-

Linguistic Feature	Label
Raw Text Properties (*RawText*)	
Sentence Length	sent_length
Word Length	char_per_tok
Vocabulary Richness (*Vocabulary*)	
Type/Token Ratio for words and lemmas	ttr_form, ttr_lemma
Morphosyntactic information (*POS*)	
Distribution of language–specific POS	xpos_dist_*
Lexical density	lexical_density
Inflectional morphology (*VerbInflection*)	
Inflectional morphology of lexical verbs and auxiliaries	verbs_*, aux_*
Verbal Predicate Structure (*VerbPredicate*)	
Distribution of verbal heads and verbal roots	verbal_head_dist, verbal_root_perc
Verb arity and distribution of verbs by arity	avg_verb_edges, verbal_arity_*
Global and Local Parsed Tree Structures (*TreeStructure*)	
Depth of the whole syntactic tree	avg_max_depth
Average length of dependency links and of the longest link	avg_links_len, max_links_len
Average length of prepositional chains and distribution by depth	avg_prep_chain_len, prep_dist_*
Clause length	avg_token_per_clause
Order of elements (*Order*)	
Relative order of subject and object	subj_pre, subj_post, obj_post, obj_pre
Syntactic Relations (*SyntacticDep*)	
Distribution of dependency relations	dep_dist_*
Use of Subordination (*Subord*)	
Distribution of subordinate clauses	subordinate_prop_dist
Average length of subordination chains and distribution by depth	avg_subord_chain_len, subordinate_dist_*
Relative order of subordinate clauses	subordinate_post

Table 1: Linguistic features used in the experiments.

tions in writing competence occurs during these two years, as shown by studies in experimental pedagogy (Barbagli et al., 2015). Accordingly, the corpus has been also successfully exploited to develop NLP-based approaches for tracking the evolution of written language competence (Miaschi et al., 2021a).

3.2 Linguistic Probing Tasks

To evaluate the linguistic competence encoded by BERT, we adopted the linguistic probing paradigm defined by Conneau et al. (2018) and followed the approach devised by Miaschi et al. (2020) who defined a suite of 68 probing tasks each corresponding to a distinct linguistic property of a sentence extracted from the IUDT sentences using Profiling-UD (Brunato et al., 2020), a tool able to acquire a wide set of linguistic features from linguistically annotated corpora according to the UD formalism. Each task consists in predicting the value that specific property has in IUDT using the representations learned by the NLM during the pre-training and fine-tuning processes. The set is reported in Table 1 and covers 9 main aspects of the structure of a sentence. They range from quite simple aspects concerning raw text properties (sentence and word length), vocabulary richness, the distribution of language-specific Parts-Of-Speech and of verbal inflectional properties (i.e. mood, tense, person).[1] More challenging probing tasks concern the ability to encode global structures of the sentence, such as the depth of the whole syntactic tree, and local features. We also paid a specific attention to testing the models knowledge of specific sub-trees, including a group of features modelling the verbal predicate structure, the order of subjects and objects with respect to their verbal head, and the use of subordination.

3.3 Models

Neural Language Model. We considered a BERT-base cased model trained on the Italian Wikipedia and texts from the OPUS corpus (Tiedemann and Nygaard, 2004).[2] Sentence-level representations are obtained using for each of the 12 layers the activation of the first input token (*[CLS]*).

Support Vector Machines Model. The SVM classifier is based on a linear kernel using the linguistic features described in Table 1. For each pair of sentences (s_i, s_j), the feature vectors V_{s_i}, V_{s_j} were combined in the final feature vector $V_{s_i,s_j} = V_{s_i} - V_{s_j}$, and normalized in the range [0,1].

Probing Model. We used a linear Support Vector Regression tested on 10,000 randomly selected IUDT sentences and trained on remaining ones. It takes as input layer-wise sentence-level representations extracted from the pre-trained and fine-tuned BERT model. As evaluation metric, we used the Spearman correlation coefficient (ρ score) between the values of the linguistic properties predicted by BERT and their gold values in IUDT sentences.

[1] For the list of Parts-Of-Speech refer to http://www.italianlp.it/docs/ISST-TANL-POStagset.pdf

[2] https://huggingface.co/dbmdz/bert-base-italian-cased

4 Predicting the Evolution of Written Language Competence

Our first experiment is aimed at assessing BERT's ability to predict the evolution of written language competence of L1 learners. Following Richter et al. (2015) and Miaschi et al. (2021a), we modelled the task as a binary classification one. We started from the assumption that, given a set of chronologically ordered essays written by the same student, a document d_j should have a higher quality level with respect to the ones written previously (d_i). However, differently from previous studies, our analysis was carried out at the sentence level. Thus, given two sentences s_i and s_j, belonging to document d_i and d_j respectively, the task consists in predicting whether $t(s_j) > t(s_i)$, where $t(s_i)$ identifies the time in which the sentence s_i was written.

To this end, we built a dataset composed of all possible pairs (s_i, s_j) of sentences contained in the *CItA* corpus and written by the same student in the two school years. We considered only the sentences contained in the first essay written during the first year and the penultimate of the second year of four different schools, for a total of 3,562 sentences and 33,566 pairs. This is motivated by the fact that this represents the widest temporal span. For each pair, we assigned two labels: 1 if $t(s_j) > t(s_i)$ and 0 otherwise. To balance the dataset, we included only one of the two labels, i.e. 50% of the pairs with label 0 (randomly selected) and the remaining of the pairs with label 1.

***CItA* statistics.** As a preliminary analysis, we investigated which linguistic phenomena are mostly involved in the writing transformations across the two years. Thus, we compared the distribution of the linguistic features automatically extracted from the set of paired sentences and described in Table 1. Table 2 shows the top 15 features ordered by decreasing number of times their value is higher in the first-year sentences.[3] The main differences concern the behaviour of verbs, both in terms of distribution of auxiliaries (*dep_dist_aux*) possibly used in compound tenses, and of characteristics specific to the verbal inflectional morphology, e.g. indicative mood (*aux_mood_dist_Ind*), and participial forms (*verbs_form_dist_Part*). The different use of connecting elements characterising the internal structure of a sentence, such as coordinating conjunctions (*xpos_dist_CC, dep_dist_cc*) and punctuation (*dep_dist_punct*), represents an additional aspect of variation.

Ranking	Feature	Difference
1	dep_dist_aux	10081
2	xpos_dist_VA	9365
3	aux_mood_dist_Ind	9046
4	verbs_form_dist_Part	8452
5	aux_form_dist_Fin	8206
6	dep_dist_conj	7642
7	aux_tense_dist_Pres	7619
8	verbs_num_pers_dist_Sing+3	-7307
9	dep_dist_cc	6493
10	xpos_dist_CC	6440
11	dep_dist_punct	-6411
12	verbs_form_dist_Fin	-5330
13	xpos_dist_E	-4707
14	verbs_tense_dist_Pres	-4688
15	xpos_dist_SP	-4525

Table 2: Ranking of first 15 features by decreasing number of times the feature value is higher in the first-year than in the second-year sentence.

4.1 Results

The SVM and BERT classification systems were evaluated with a 4-fold cross-validation in which every fold uses as a test set the sentences of the school not included in the corresponding training set. The accuracy for each fold refers to the percentage of sentences correctly classified and the final accuracy is the average score.

Figure 1 reports the task performances achieved by BERT for each of the 50 fine-tuning epochs. As we can see, the highest accuracy is 0.78 and it is achieved at the 11th epoch. It is worth noting that lower accuracy, ranging from 0.73 to 0.76, was obtained at epochs 2, 3 and 4 which were suggested as optimal hyperparameters in many tasks by Devlin et al. (2019).

Note that BERT outperforms the results achieved by the SVM classifier, which is able to solve the task with an accuracy of 0.70. The relatively small difference in terms of accuracy between a deep NLM and a simple linear classifier, which does not exploit words of the sentence as features but only linguistic properties, seems to confirm our intuition regarding the linguistic nature of the chosen task. A more in-depth analysis focused on the contribution of specific linguistic features used by the SVM classifier highlights that the most predicting ones are represented by the use of verbs, in particular of the auxiliary verbs, and of punctuation and conjunctions. This is in line with the statistics reported

[3] All variations across the two years are statistically significant according to the Wilcoxon Rank Sum Test.

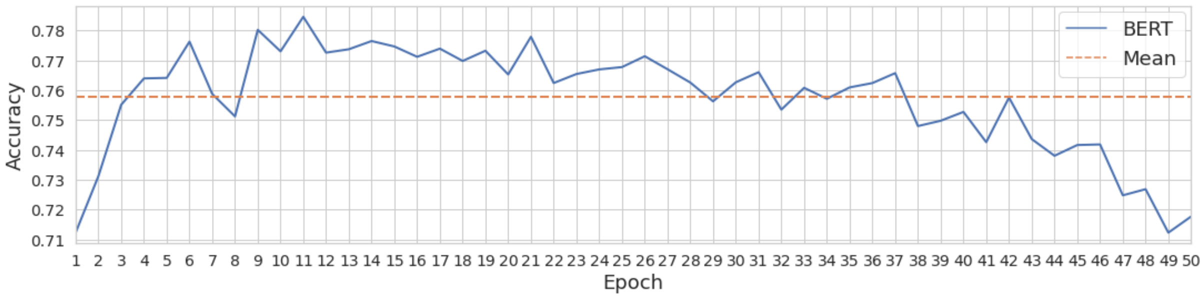


Figure 1: BERT accuracy on the downstream task across 50 fine-tuning epochs (*BERT*) and on average (*Mean*).

in Table 2, where we showed that a change in the distribution of these linguistic features is among the most relevant writing transformations occurring across the two school years.

5 How Does Linguistic Competence Change Across the Fine-tuning Epochs?

BERT linguistic competence was tested by adopting a supervised probing classifier approach, which accounts for 68 probing tasks. As detailed in Section 3.3, we trained a LinearSVR probing model which uses the sentence-level representations extracted from the pre-trained and fine-tuned model. Since many of our probing features are strongly related to sentence length, we compared the results with the ones obtained using a baseline computed by using a probing model trained using only sentence length as an input feature.

Pre-training linguistic competence. Figure 2 reports the probing task results obtained using the pre-trained BERT's representations. As a general remark, we can notice that BERT's scores always outperform the baseline ones, with the obvious exception of the sentence length (*n_tokens*). However, as expected, the results are more similar for the probing tasks whose resolution is highly influenced by the length of the sentence. This is the case of global and local properties of the sentence such as the depth of the syntactic tree (*avg_max_depth*) and the length of the longest dependency links (*max_links_len*). They correspond to linguistic phenomena that the NLM masters very well. Nevertheless, differently from the baseline, BERT shows to encode quite accurately also the only lexical property we considered, i.e. the Type/Token ratio, and a raw text feature highly related to the use of lexicon, i.e. the length of tokens (*char_per_tok*). Pre-trained representations are also very good at predicting the distribution of functional information both in terms of Parts-Of-Speech such as prepositions (*xpos_dist_E*), coordinating conjunctions (**_CC*), determinative articles (**_RD*), and of dependency relations linking a functional POS to its head (i.e. *dep_dist_case*, **_cc*, **_det* respectively), and of punctuation, in particular of commas (**_FF*) and full stops (**_FS*). Lastly, concerning BERT's knowledge about verbs, we can note that the model encodes quite accurately the tense, person and mood of auxiliary verbs (*aux_**) and the distribution of verbal heads (*verbal_head_per_sent*).

When we move to the analysis of how the competence changes across layers, we can observe that it generally tends to decrease when the output layer is approaching. This is in line with previous findings (Liu et al., 2019; Miaschi et al., 2020) and it could be due to the fact that Transformer layers trade off between task-oriented (e.g. Masked Language Modeling) information and general linguistic competence. A such decreasing trend can be specifically appreciated by observing the evolution of the black dotted line reported in Figure 3. It shows how BERT's competence in the 9 groups of linguistic phenomena introduced in Table 1 evolves across the 12 layers. As it can be seen, however not all features have the same decreasing trend. The main exceptions are represented by quite complex linguistic features, such as the order of subject/object with respect to the verbal head (*Order of elements*), the use of subordination and the verbal predicate structure, which increase consistently across layers, even if they decrease in the output layer. These linguistic features model syntactic sub-trees of the sentence, that require an in-depth knowledge of the syntactic structure of the sentence, which is encoded starting from the middle layers, as shown for English (Tenney et al., 2019; Miaschi et al., 2020).

Fine-tuning impact. Coloured lines in Figure 3 report the layer-wise ρ scores we obtained us-

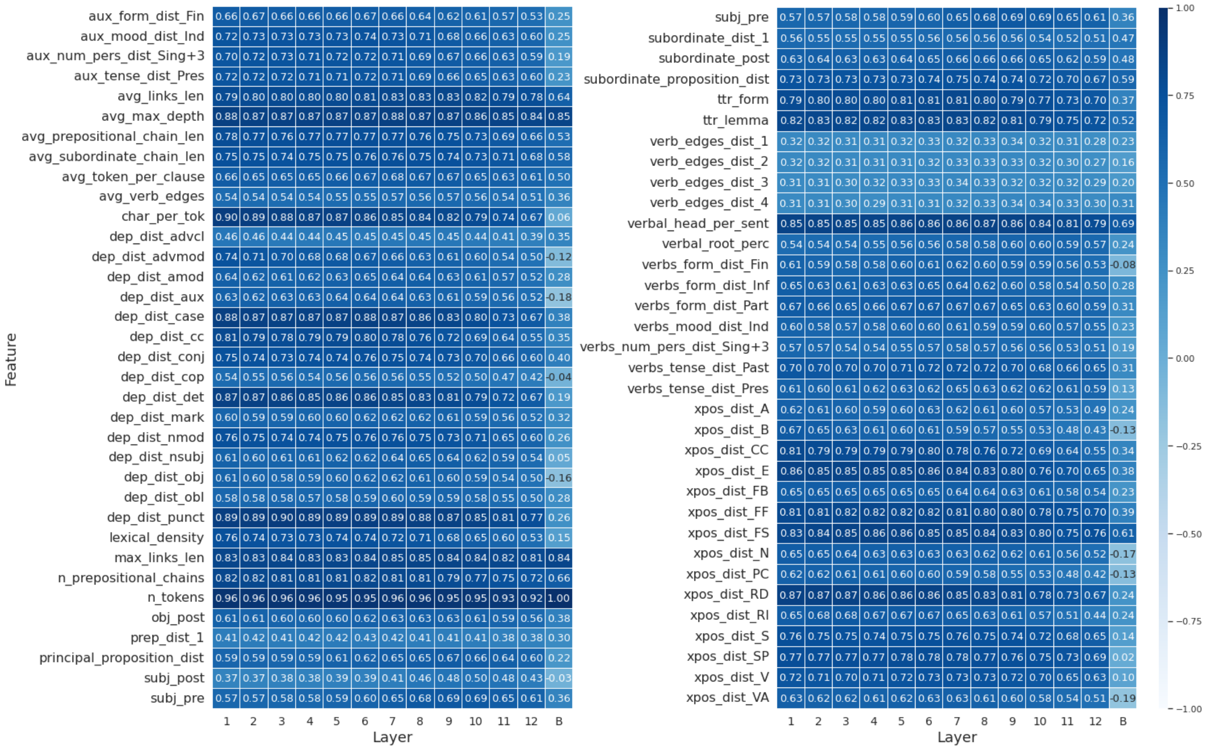


Figure 2: Layer-wise ρ scores for the 68 probing tasks obtained with the pre-trained model. Baseline score are reported in column B.

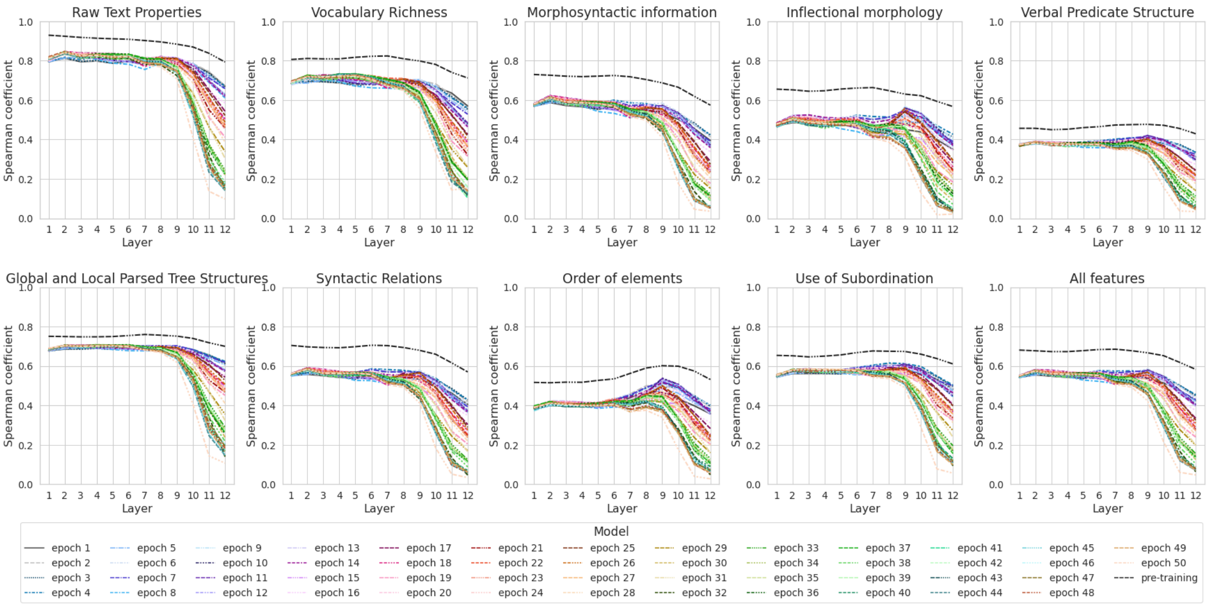


Figure 3: Pre-trained and fine-tuned BERT layerwise ρ scores for groups of probing features.

ing the representations extracted by each of the 50 epochs of fine-tuning on the *CItA* corpus. Differently from scholars who found that some linguistic abilities are reinforced after the fine-tuning process, such as Merchant et al. (2020) and Mosbach et al. (2020), but more similarly to Miaschi et al. (2020), our analysis shows that BERT tends to lose the linguistic competence acquired during the pre-training phase. However, such loss has a different impact across the 50 epochs and the 12 layers. For all epochs, all linguistic competencies are mostly altered starting from the 8th layer, thus resulting mostly modified in the 12th one. This might be possibly due to the fact that during these layers the

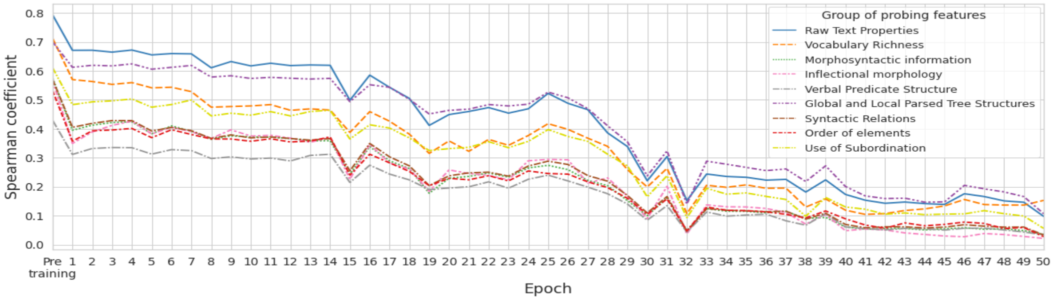


Figure 4: ρ scores for groups of probing features at the 12th layer of pre-training and across the 50 epochs of fine-tuning.

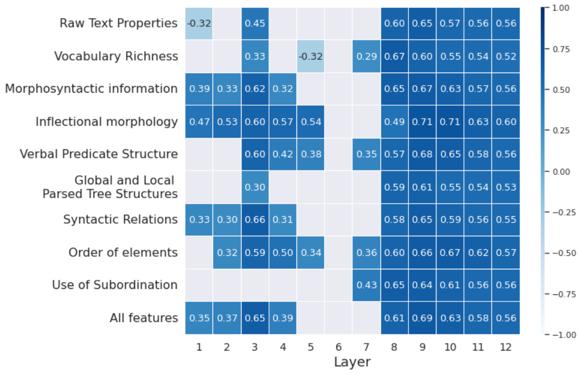
model is storing task–specific information at the expense of its ability to encode general knowledge about the language (Kovaleva et al., 2019).

In addition, note that the impact of the fine-tuning stage is less evident in the first epochs (blue lines in the figure) than in the last ones (green lines). Since the main changes are observed in the 12th layer, we deepen the analysis and report in Figure 4 the probing scores for each group of features at each epoch. The trend shows that for all groups of features a first main drop (with respect to the pre-training scores) occurs after the first epoch, then the sentence properties are constantly forgotten across the 50 epochs. However, such a negative impact is different for each group. For example, the knowledge of simple raw text features, which BERT knows very well, drops from 0.8 to 0.1. Similarly, the generalization abilities of a highly related group of features, i.e. those referring to global and local syntactic characteristics of a sentence, decrease quite considerably.

Groups of probing features	slope	r-value
Raw Text Properties	1.679	0.966
Global and Local Parsed Tree Structures	1.644	0.955
Use of Subordination	1.622	0.961
Verbal Predicate Structure	1.618	0.964
Inflectional morphology	1.581	0.955
All features	**1.574**	**0.965**
Syntactic Relations	1.542	0.965
Order of elements	1.534	0.964
Morphosyntactic information	1.534	0.965
Vocabulary Richness	1.449	0.958

Table 3: Ranking of groups of probing features according to decreasing slope of the regression lines. Correlation coefficients are also reported (*r-value*).

To further investigate which group of features is forgotten most rapidly across the epochs, we calculated a slope of a linear regression line between the 50 epochs and the decay of competencies for each linguistic group.[4] Results are reported in Table 3 where we show the 9 groups of features ordered according to decreasing slope values. As it can be noted, the ranking does not reflect BERT's linguistic competence. Namely, in the top and last position we find groups of features that pre-trained BERT knows very well, i.e. Raw text properties and Vocabulary Richness. Similarly, the groups that are scarcely mastered are scattered along the ranking. This might suggest that the change of competencies across the epochs is not related to the pre-trained knowledge but possibly to some characteristics useful in the resolution of the downstream task.



	1	2	3	4	5	6	7	8	9	10	11	12
Raw Text Properties	-0.32		0.45					0.60	0.65	0.57	0.56	0.56
Vocabulary Richness			0.33		-0.32		0.29	0.67	0.60	0.55	0.54	0.52
Morphosyntactic information	0.39	0.33	0.62	0.32				0.65	0.67	0.63	0.57	0.56
Inflectional morphology	0.47	0.53	0.60	0.57	0.54			0.49	0.71	0.71	0.63	0.60
Verbal Predicate Structure			0.60	0.42	0.38		0.35	0.57	0.68	0.65	0.58	0.56
Global and Local Parsed Tree Structures			0.30					0.59	0.61	0.55	0.54	0.53
Syntactic Relations	0.33	0.30	0.66	0.31				0.58	0.65	0.59	0.56	0.55
Order of elements		0.32	0.59	0.50	0.34		0.36	0.60	0.66	0.67	0.62	0.57
Use of Subordination							0.43	0.65	0.64	0.61	0.56	0.56
All features	0.35	0.37	0.65	0.39				0.61	0.69	0.63	0.58	0.56

Figure 5: Layer-wise correlation between the average ρ score for groups of features and BERT's accuracy on the downstream task. Blank cells are non statistically significant correlations.

6 Task Resolution and Linguistic Competence

Correlating task accuracy with linguistic competence. For each single feature and each group of linguistic features, Figures 5 and 6 report the results of

[4] decay = (probing score at layer 12 of pre-training - probing score at layer 12 epoch *i*) / probing score at 12 layer of pre-training

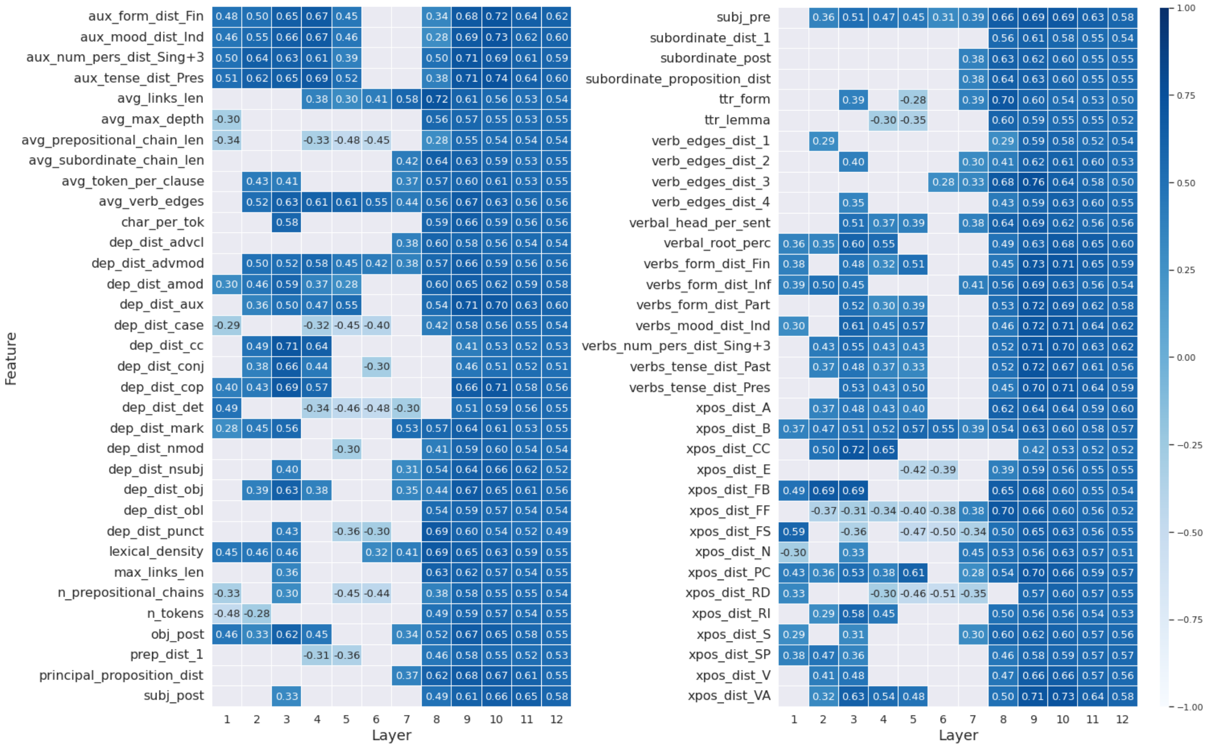


Figure 6: Layer-wise correlation between the Spearman coefficient value of the 68 linguistic features and the accuracy over the 50 epochs considered for the fine-tuning.

the Spearman correlations between the layer-wise probing scores and the accuracy of BERT in the resolution of the downstream task for all epochs. The results were compared with a baseline computed as the Spearman correlation between BERT's accuracy on the downstream task and a list of numbers from 50 to 1 (ρ=0.55)[5]. As a general remark, in the first layers (from 1st to 7th), we noticed a non-significant or low correlation, most of the time lower than the baseline scores. On the contrary, the scores are higher, always above the baseline, in the central layers (8-10) and they tend to decrease in the 11th and 12th layers, where task-specific information is stored. The trend suggests that the higher BERT's linguistic competence, the more accurate the resolution of the downstream task is. This is a global trend as showed by *All features*, even if there are some differences among the studied linguistic phenomena. Namely, information about verbs (both in terms of inflectional morphology and predicate structure) and the order of nuclear elements of sentence results to be the linguistic competence mostly correlated with the resolution of the task. In particular, the distribution of auxiliary verbs (both in terms of POS and dependency relations), clitic pronouns, tense, person and mood of both auxiliary and lexical verbs, of subjects in their canonical order (i.e. in the pre-verbal position), show a very high correlation ($\rho \geq 0.70$) in either the 9th or 10th layer. Interestingly, these linguistic phenomena are the mostly involved in the writing transformations across the two school years (see Table 2).

Predicting task accuracy with linguistic competence. The last experiment is aimed at testing whether it is possible to predict BERT's accuracy in the resolution of the downstream task on the basis of its linguistic competence assessed by the linguistic probing tasks. For this purpose, we trained a linear Support Vector Regressor using a Leave-one-out strategy. Namely, each time, we used as training the average accuracy of the probing tasks (in terms of ρ scores) achieved at the 12th layer of 49 epochs, and the excluded epoch as the test. As an evaluation metric, we used the Spearman correlation and the Mean Squared Error (MSE) between the predicted values and the actual ones. We devised two baselines: the first one consists of a list of 50 numbers corresponding to the average BERT's

[5] Such baseline was chosen since it simulates the decreasing trend of probing score accuracy during the 50 epochs of fine-tuning as shown in Figure 4.

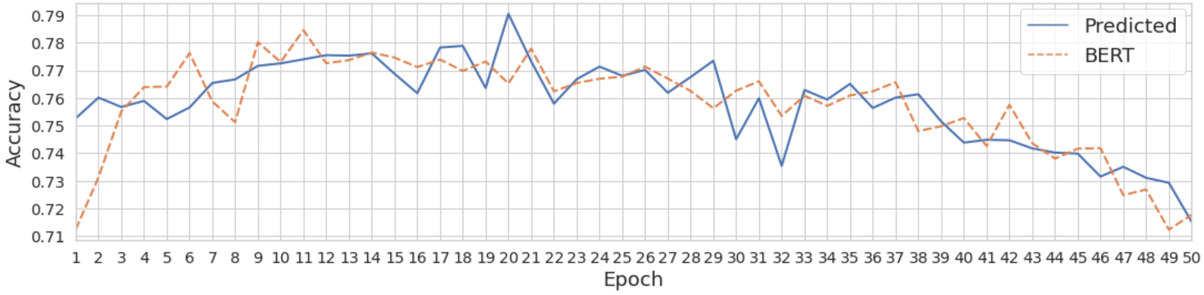


Figure 7: Predicted accuracy at the 12th layer and BERT's actual accuracy for each epoch of fine-tuning.

accuracy in the resolution of the downstream task considering all the epochs. In this case, since the value is always the same for all the epochs, the prediction was evaluated only in terms of MSE. The second baseline is a linear SVR using a decreasing list of numbers from 50 to 1, simulating the decreasing trend of the probing accuracy across the epochs. Here, we were able to compute also the correlation scores. The results show that the SVR achieves very high scores (ρ=0.76, MSE=0.011) which outperform the two baselines (ρ=NA, MSE=0.017; ρ=0.54, MSE=0.017). This demonstrates that the BERT accuracy can be reliably predicted by studying its linguistic abilities. The reliability of this result is corroborated by Figure 7 where we compare the trends of the predicted (*Predicted*) and BERT actual accuracy. It is worth noting that the two lines are almost overlapping, showing a very similar trend.

7 Conclusion

In this paper, we carried out an extensive study aimed at investigating in detail the relation between the fine-tuning stage and the linguistic knowledge encoded by BERT. For our study, we chose the prediction of the evolution of written language competence of native language learners as a downstream task, since the morpho-syntactic and syntactic information plays an important role in the resolution. We showed in particular how the knowledge assessed during the pre-training stage is progressively forgotten across 50 epochs of fine-tuning. We observed that the main changes concern the 12th layer of all epochs and they are more evident in the last than in the first epochs. An in-depth analysis showed that the types of linguistic phenomena forgotten most rapidly across epochs are not related to the linguistic knowledge of the pre-trained NLM but possibly to some characteristics useful to solve the downstream task. We also found a strong correlation between the linguistic knowledge encoded by BERT and its ability to resolve the task. Specifically, we showed that the verbal inflectional morphology, the predicate structure, and the canonical order of subject and object are the most correlated aspects. Interestingly, these linguistic characteristics, in particular those referring to verbs, resulted to be the most predictive types of features exploited by the SVM classifier and the most involved in the writing transformations. A such strong relationship between the ability to the resolution of the task and linguistic competence is particularly evident in our last experiment, where we obtained very strong accuracy in predicting BERT's performance using the encoded linguistic competence. Besides investigating the nature of BERT this outcome could open the way to define a new linguistically motivated stop-function of the fine-tuning process.

Ethical Considerations

The findings described in this work are based on the analysis of essays written by Italian L1 learners and they are mostly intended to evaluate recent efforts in the interpretability of neural language models. As we used authentic texts from students, we took care of anonymization separating their personally identifying information from their essays to prevent any harmful uses of data.

Acknowledgements

We acknowledge the project "Human in Neural Language Models" (*IsC93_HiNLM*), funded by the CINECA[6] under the ISCRA initiative, for the availability of HPC resources and support.

References

Alessia Barbagli, Pietro Lucisano, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2015. Il

[6] https://www.cineca.it/en

ruolo delle tecnologie del linguaggio nel monitoraggio dell'evoluzione delle abilità di scrittura: primi risultati. *Italian Journal of Computational Linguistics (IJCoL)*, vol. 1(1):99–117.

Alessia Barbagli, Pietro Lucisano, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2016. CItA: an L1 Italian learners corpus to study the development of writing competence. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 88–95, Portorož, Slovenia. European Language Resources Association (ELRA).

Yonatan Belinkov. 2021. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, pages 1–13.

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

Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2020. Profiling-ud: a tool for linguistic profiling of texts. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7147–7153, Marseille, France. European Language Resources Association.

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

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

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

Nadir Durrani, Hassan Sajjad, and Fahim Dalvi. 2021. How transfer learning impacts linguistic knowledge in deep NLP models? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4947–4957, Online. Association for Computational Linguistics.

Allyson Ettinger. 2020. What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Raffaele Guarasci, Stefano Silvestri, Giuseppe De Pietro, Hamido Fujita, and Massimo Esposito. 2022. Bert syntactic transfer: A computational experiment on italian, french and english languages. *Computer Speech & Language*, 71:101261.

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.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4365–4374, Hong Kong, China. Association for Computational Linguistics.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. 2020. What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 33–44, Online. Association for Computational Linguistics.

Alessio Miaschi, Dominique Brunato, and Felice Dell'Orletta. 2021a. A NLP-based stylometric approach for tracking the evolution of L1 written language competence. *Journal of Writing Research*, vol. 13(1):71–105.

Alessio Miaschi, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2020. Linguistic profiling of a neural language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 745–756, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alessio Miaschi, Gabriele Sarti, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2021b. Italian transformers under the linguistic lens. In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, Online. CEUR Workshop Proceedings (CEUR-WS.org).

Marius Mosbach, Anna Khokhlova, Michael A. Hedderich, and Dietrich Klakow. 2020. On the interplay between fine-tuning and sentence-level probing for linguistic knowledge in pre-trained transformers. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 68–82, Online. Association for Computational Linguistics.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377, Online. Association for Computational Linguistics.

Stefan Richter, Andrea Cimino, Felice Dell'Orletta, and Giulia Venturi. 2015. Tracking the evolution of written language competence: an nlp–based approach. *CLiC it*, page 236.

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

Gabriele Sarti, Dominique Brunato, and Felice Dell'Orletta. 2021. That looks hard: Characterizing linguistic complexity in humans and language models. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 48–60, Online. Association for Computational Linguistics.

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.

Jörg Tiedemann and Lars Nygaard. 2004. The opus corpus-parallel and free: http://logos. uio. no/opus. Citeseer.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretic, and Samuel R. Bowman. 2019. Investigating BERT's knowledge of language: Five analysis methods with NPIs. 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 2877–2887, Hong Kong, China. Association for Computational Linguistics.

Lang Yu and Allyson Ettinger. 2021. On the interplay between fine-tuning and composition in transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2279–2293, Online. Association for Computational Linguistics.