MDPI

Article

Testing the Effectiveness of the Diagnostic Probing Paradigm on Italian Treebanks

Alessio Miaschi [1,*], Chiara Alzetta [1], Dominique Brunato [1], Felice Dell'Orletta [1] and Giulia Venturi [1]

[1] Institute for Computational Linguistics "A. Zampolli" (ILC-CNR), ItaliaNLPLab, via G. Moruzzi 1, Pisa, Italy; name.surname@ilc.cnr.it
* Correspondence: alessio.miaschi@ilc.cnr.it

Abstract: The outstanding performance recently reached by Neural Language Models (NLMs) across many Natural Language Processing (NLP) tasks has fostered the debate towards understanding whether NLMs implicitly learn linguistic competence. Probes, i.e. supervised models trained using NLM representations to predict linguistic properties, are frequently adopted to investigate this issue. However, it is still questioned if probing classification tasks really enable such investigation or if they simply hint at surface patterns in the data. This work contributes to such debate by presenting an approach to assess the effectiveness of a suite of probing tasks aimed at testing the linguistic knowledge implicitly encoded by one of the most prominent NLMs, BERT. To this aim, we compared the performance of probes when predicting gold and automatically altered values of a set of linguistic features. Our experiments, performed on Italian, extend the work of Miaschi *et al.* [1] evaluating the results across BERT layers and for sentences with different lengths. As a general result, we observed higher performance in the prediction of gold values, thus suggesting that the probing model is sensitive to the distortion of feature values. However, our experiments also showed that the length of the sentence is a highly influential factor that is able to confound the probing model's predictions.

Keywords: neural language models; bert; probing tasks; treebanks; italian language

Citation: Lastname, F.; Lastname, F.; Lastname, F. Title. *Information* **2022**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

1. Introduction

The rise of large pre-trained Neural Language Models (NLMs) has revolutionized the field of Natural Language Processing (NLP) in the last five years. In particular, the introduction of deep contextualized models based on the Transformer architecture [2], able to learn word vectors that are sensitive to the context in which words appear, has yielded significant improvements on many NLP tasks [3–5]. Even with some differences concerning the size of their parameters, architectures, and training datasets [6–8], these models are all pre-trained on large amounts of text and subsequently fine-tuned on task-specific, supervised downstream tasks. Among the many Transformer-based models, BERT (*Bidirectional Encoder Representations from Transformers*) has been the first one to push the state of the art in many areas of NLP [9].

However, it is well known in the literature that the remarkable ability of BERT, and NLMs in general, to perform numerous language-understanding tasks goes with an opaqueness concerning the interpretation of their internal mechanisms. Particular interest has been devoted in the last few years to the investigation of the linguistic abilities implicitly encoded by the models [10]. Namely, several methods have been proposed to obtain meaningful explanations of how NLMs are able to capture syntax- and semantic-sensitive phenomena [11], also taking inspiration from human language experiments [12,13]. They range from the analysis of attention mechanisms [14] and the definition of diagnostic tests [15] to the implementation of explainability techniques via e.g. integrated gradients [16]. One of the most 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 [17].

However, despite the amount of work based on the diagnostic probing approach, as outlined by Belinkov [18] there are still several open questions, such as: which probing model should we use for assessing the linguistic competence of a NLM? Are probes the most effective strategy to achieve such a goal? These questions fostered two complementary lines of research. The first one is devoted to modifying the architecture of the current probing models; the other one is focused on evaluating their effectiveness. Both are still not well-investigated issues, although their importance for advancing the research on the evaluation of NLMs linguistic competencies has been widely recognized.

This study would contribute to the debate on the effectiveness of the probing paradigm as a diagnostic method to assess the linguistic knowledge implicitly encoded by BERT. To achieve this goal, we define a multifaced approach that comprises a number of experiments aimed at comparing the performance of a probing model trained using BERT representations to predict the values of a set of sentence-level properties extracted from the Italian Universal Dependency Treebank [19] and from a suite of *control datasets* that we specifically built for the purpose of this study. Starting with and extending the methodology introduced by Miaschi *et al.* [1], we define as control dataset a set of linguistic features whose values are automatically altered in order to be increasingly different from the values in the treebank referred to as *gold* values. Our underlying hypothesis is the following: if the probing model's predictions of the variously altered values diverge from the predictions of the gold values, this possibly suggests that the corresponding probing tasks are effective strategies to test the linguistic knowledge embedded in BERT representations. We will discuss the results of the experiments in light of this hypothesis. The remainder of the paper is organised as follows. We present our background and related work in Section 2. Section 3 introduces our methodology, presenting the data, the monitored linguistic features and the models used in the study. Section 4 presents the results and in Section 5 we will draw the conclusions.

Contributions

With respect to previous literature, the main contributions of our work lie in the following points:

- We present a methodology to test the reliability of probing tasks by building control datasets at diverse levels of complexity;
- We assess to which extent the linguistic knowledge encoded by BERT is influenced by the length of the sentence and how the length can represent a confounding factor that may bias the real estimate of BERT's knowledge of a wide variety of (morpho)-syntactic phenomena;
- We test the effectiveness of the diagnostic probing task approach on Italian, a language frequently neglected by studies on probing.

2. The Diagnostic Probing Paradigm

In the last few years, the analysis of the inner workings of state-of-the-art Neural Language Models (NLMs) has become one of the most popular lines of research in NLP. In particular, great efforts have been devoted to obtaining meaningful explanations about their linguistic competence in order to understand to what extent these models are able to capture linguistic properties targeting a variety of domains [20]. These approaches range from the definition of fill-the-gap probes [15] and probing tasks that a model can only solve if it has encoded a precise linguistic phenomenon [17,21,22], to the analysis of attention mechanism [23–25] and correlations between representations [26].

Among the different strategies developed to study the implicit language competencies encoded by NLMs, the *diagnostic probing task* approach has emerged as one of the most commonly adopted ones. The idea behind the probing paradigm is actually quite simple: using a diagnostic classifier, the *probing model* or *probe*, which takes the output representations of a NLM as input, to perform a *probing task*, e.g. predict a given language

property. If the probing model will predict the property correctly, then we can assume that the representations somehow encode that property.

Studies relying on this approach reported that NLMs contextual representations are able to encode a broad spectrum of linguistic properties, from information about Parts-of-Speech (POS) and other morphological properties to syntactic and semantic information. In particular, these works demonstrated that NLMs learn a variety of language properties in a hierarchical manner [11,27,28] and that their representations also support the extraction of dependency parse trees [29]. Training a simple probing classifier that has access only to the per-token contextual embeddings of a BERT model, Tenney *et al.* [30] showed that the order in which specific abstractions are encoded within the internal representations reflects the traditional hierarchy of the NLP pipeline: POS tags are processed earliest, followed by constituents, dependencies, semantic roles, and coreference. Liu *et al.* [31], instead, quantified differences in the transferability of individual layers between different NLMs, showing that higher layers of ELMo [32] are more task-specific (less general), while transformer layers (BERT) do not exhibit this increase in task-specificity.

Despite this emerging body of work, there are still several open questions about how probing tasks should be designed, how complex a probe should be allowed to be, or whether probes are actually showing the linguistic generalization abilities of the NLMs rather than learning the linguistic tasks [18]. Among the first line of research, which deals with the design of probing classifiers, several works investigate which model should be used as probe and which metric should be employed to measure their performance. With this respect, it is still questioned if one should rely on simple models [29,31,33] or more complex ones [34,35] in terms of model parametrization. For instance, Voita and Titov [35] suggest designing alternative probes using a novel information-theoretic approach which balances the probe's inner complexity with its task performance. Although this line of research raises many interesting questions, in this work we take the distance from it and investigate the probing paradigm from a different viewpoint.

Our perspective is closer to the second line of research on the probing task approach, which indeed is concerned with testing the evaluation of the effectiveness of probing models. Embracing such a line, for example, Hewitt and Liang [21] suggested that probing tasks might conceal the information about the NLM representation behind the ability of the probe to learn surface patterns in the data. To test this intuition, they introduced the idea of *control tasks*, a set of tasks that associate word types with random outputs that can be solved by simply learning regularities. Measuring the difference between the accuracy on linguistic tasks and on control tasks (a property defined as *selectivity*) they identified 'good' probes as the ones for which the model achieves high linguistic task accuracy and low control task accuracy, thus providing insights into the linguistic properties of a representation. Along the same line, Ravichander *et al.* [36] test probing tasks by creating control datasets where a property is always reported in a dataset with the same value, thus it is not discriminative for testing the information contained in the representations. Their experiments highlight that the probe may learn a property also incidentally, thus casting doubts on the effectiveness of probing tasks.

While sharing the same goal as these previous works, our study differs in two main respects. Firstly, we follow an approach similar to Hewitt and Liang [21] but we introduce a methodology to progressively test the effectiveness of probing models, by devising diverse control tasks differing at the level of increasing complexity and which intend to address a larger set of linguistic phenomena. Secondly, we focus on the Italian language, which is much less explored in the area of interpretability. In fact, the majority of research is focused on English or, at most, multilingual models, with only a few exceptions [37–39].

3. Methodology

The methodology we devised is aimed at testing whether a diagnostic probing model really encodes the linguistic competencies of a NLM or simply learns the regularities of one or more probing tasks. To this aim, we trained a probing model using BERT

sentence representations, as described in Section 3.4, and then tested its performance in the resolution of a set of linguistic tasks. These tasks consist in predicting the values of various linguistic features (see Section 3.2) extracted from different sections of the Italian Universal Dependency Treebank (IUDT).

The probing model was tested in two main scenarios. In the first one, the model has to predict gold features, i.e. the real values of the features in IUDT sentences. In the second scenario, gold values have been altered based on multiple strategies in order to obtain alternative datasets at different control levels. As discussed in Section 3.3, this scenario, which is articulated into multiple ones, is based on the rationale that if the predictions of the probing model are more accurate and thus more similar to the gold values than to the automatically altered ones, then we might assume that BERT's representations do encode the linguistic knowledge required to solve the task. Consequently, the intuition is that the probing model has not simply learned some regularities possibly found in the dataset and used them to solve the linguistic task.

3.1. Data

For our experiments, we relied on the Italian Universal Dependencies Treebank (IUDT), version 2.5. IUDT contains a total of 35,480 sentences and 811,488 tokens, and it consists of a combination of four sections, representative of the standard Italian language, i.e. the Italian version of the multilingual Turin University Parallel Treebank (ParTUT) [40], the Venice Italian Treebank (VIT) [41], the Italian Stanford Dependency Treebank (ISDT) [42], PUD [43], and of two sections including examples of social media texts, i.e. PoSTWITA [44], TWITTIRÒ [45].

Considering the high variability in terms of sentence length in IUDT, which contains sentences ranging from 1 to 310 tokens long, we decided to split the treebank into three subsets, containing respectively the **shortest**, the **standard** and the **longest** sentences. The larger subset is the *Standard* one: it contains 21,991 sentences having a length between 10 and 30 tokens. This is a quite typical length in Italian, a language in which the average sentence length is equal to about 20 tokens, like in this example sentence acquired from the Standard subset '*Un rumore infernale, simile al passaggio di un treno, risuona nei corridoi sotterranei che solcano Rochester*' (trad. 'An infernal noise, similar to the passage of a train, resounds in the underground corridors that run through Rochester').

The other two subsets comprise sentences whose length is less standard. Within the *Shortest* subset, we included 5,538 sentences whose length is up to 9 tokens. This set covers many examples of nominal or elliptical sentences, including for instance news titles (e.g. *'Battesimo per l'opera verdiana.'*, trad. 'Baptism for Verdi's opera.'), short questions (e.g. *'Come si spiega un simile risultato?'*, trad. 'How can such a result be explained?') and sentences showing a quite simple syntactic structure (e.g. *'Questa ricchezza è tutta apparenza.'*, trad. 'This wealth is all appearance'). Note that we excluded for this subset sentences having less than 3 tokens (288 in the dataset) since they do not show a proper syntactic structure given that they generally consist of a single token plus punctuation. The set of long sentences, on the other hand, comprises sentences whose length ranges between 31 and 100 tokens and it contains 7,585 sentences. The following 58 tokens-long sentence represents a quite typical example of sentences belonging to the Longest subset: *'Una giornata convulsa durante la quale il presidente della Regione Lazio, Renata Polverini, è arrivata vicina alle dimissioni in seguito alla crisi generata dall'abuso di fondi pubblici da parte del Pdl laziale per il quale è indagato, con l'accusa di peculato, l'ex capogruppo Franco Fiorito.'*, trad. 'It was a convulsive day during which the President of the Lazio Region, Renata Polverini, came close to resigning following the crisis generated by the misuse of public funds by the Lazio PDL, for which former group leader Franco Fiorito is under investigation on charges of embezzlement.'. IUDT reports additional 78 sentences longer than 100 tokens, which we excluded from the experiments since we noticed that they are characterized by a debatable annotation, possibly caused by an erroneous sentence splitting. Note that, for the specific purpose of the experiments

Table 1. Probing Features used in the experiments grouped into 7 main types of linguistic phenomena.

Linguistic Feature	Label
Order of elements (*Order*)	
Relative order of subject and object	subj_pre, subj_post, obj_post
Morphosyntactic information (*POS*)	
Distribution of UD and language-specific POS	upos_dist_*, xpos_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_1
Relative order of subordinate clauses	subordinate_post
Syntactic Relations (*SyntacticDep*)	
Distribution of dependency relations	dep_dist_*
Global and Local Parsed Tree Structures (*TreeStructure*)	
Depth of the whole syntactic tree	parse_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_1
Clause length	avg_token_per_clause
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_*

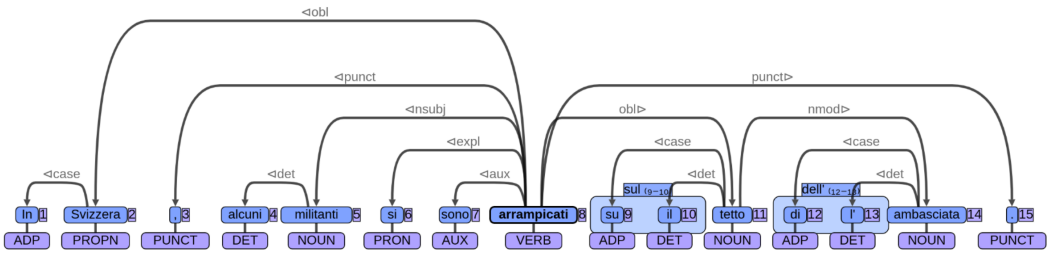

Figure 1. Linguistic annotation based on the UD scheme of the example sentence.

conducted in this study, we undersampled the Longest set to 5,538 sentences, which we randomly selected, in order to balance it to the set of sentences in the Shortest subset.

3.2. Linguistic Features

In order to probe the linguistic competence encoded by the language model, we relied on the approach for the first time proposed by Miaschi *et al.* [46] which consists in predicting the value of multiple linguistic features of a sentence using the model's representations. The set of linguistic features is based on the one described in Brunato *et al.* [47] that includes about 130 features representative of the linguistic structure underlying a sentence and derived from raw, morpho-syntactic and syntactic levels of annotation. In this study, we selected the 77 most frequent features occurring in the IUDT sections in order to prevent data sparsity issues. As can be seen in Table 1, they are grouped into seven main types of linguistic phenomena which range from morpho-syntactic and inflectional properties to more complex aspects of sentence structure (e.g. the depth of the whole syntactic tree), to features referring to the structure of specific sub-trees, such as the relative order of subjects and objects with respect to the verb, to the use of subordination.

We chose to rely on these features for two main reasons. Firstly, they have been shown to be highly predictive when leveraged by traditional learning models on various classification problems where linguistic information plays a fundamental role [47]. In addition, they are multilingual as they are based on the Universal Dependency formalism for sentence representation [48]. In fact, they have been successfully used to profile the knowledge encoded in the language representations of contextual NLMs for both the Italian [38] and English language [22].

Figure 1 exemplifies some of them extracted from the following sentence acquired from the Standard subset:

(1) *In Svizzera, alcuni militanti si sono arrampicati sul tetto dell'ambasciata.* [trad. 'In Switzerland, some militants climbed onto the roof of the embassy.']

Table 2. Average values and coefficient of variation of each macro-group of *gold* linguistic features, extracted from the sentences in the Shortest, Standard and Longest subsets of IUDT.

	Shortest		**Standard**		**Longest**	
Feat. Group	*Mean*	*CV*	*Mean*	*CV*	*Mean*	*CV*
Order	19.83	1.04	40.45	0.55	52.96	0.34
POS	3.56	0.14	3.56	0.09	3.68	0.03
Subord	16.73	0.96	36.51	0.58	48.67	0.29
SyntacticDep	5.36	0.19	5.33	0.13	5.51	0.08
TreeStructure	4.62	1.17	11.66	0.56	17.37	0.29
VerbInflection	23.21	0.80	38.38	0.47	47.38	0.33
VerbPredicate	16.60	0.78	23.17	0.39	25.97	0.22

Relying on the morpho-syntactic level of IUDT annotation, we can observe for example that the above sentence features 20% of prepositions (ADP), 6.66% of verbs (VERB), and 20% of nouns (NOUN) out of the total amount of Parts-Of-Speech. Considering the features referring to the global syntactic structure, the depth of the whole syntactic tree of the sentence is equal to 3, corresponding to the two intermediate dependency links that are crossed in the path going from the root of the sentence (*arrampicati*, 'climbed') to each of the more distant leaf nodes, represented by the words *di* ('of') and *l'* ('the'), which compose the articulated preposition *dell'* dependent of the word *ambasciata* ('embassy'). Focusing on the local tree structure, the longest dependency relation is 6-token long, which corresponds to the number of tokens occurring linearly between the syntactic head *arrampicati* ('climbed') and the oblique object (obl) *Svizzera* ('Switzerland'), and we can observe a 1-link long prepositional complement chain (nmod) *dell'ambasciata* ('of the embassy') headed by the noun *tetto* ('roof'). Besides, the sentence is characterized by a canonical order of nuclear elements since the nominal subject *militanti* ('militants') is in a pre-verbal position, which is the preferred order in Italian.

In this study, the values of each feature acquired from IUDT represent the *gold values*. Table 2 reports the average distribution (*Mean*) and coefficient of variation (*CV*) of each group of linguistic features, computed as a mean of the values of every single feature included in the group. As it can be noted, the Mean values vary consistently across the three IUDT subsets since we account for many different linguistic phenomena characterized by diverse ranges of values. As expected, most features are influenced by the length of the sentences being considered. In fact, while the mean values increase along with sentence length, the coefficients of variation, which capture the extent of values variability within the same subset, tend to decrease as we approach the Longest subset. This suggests that, as sentences get longer, linguistic features tend to show higher but more stable values, while the opposite happens on sentences belonging to the Shortest subset. For the purposes of our experiments, the *gold values* reported in the gold dataset (IUDT) have been automatically altered to generate *control datasets*.

3.3. Control datasets

We created two main types of control datasets for each subset of IUDT, obtained by automatically altering gold feature values according to different strategies. The first main type, hereafter referred to as *Swapped*, is built by shuffling the original values of each feature across sentences; the second type, *Random*, contains values randomly generated within the maximum and the minimum value that each feature shows in the gold datasets. To clarify, consider the following example involving the feature `average link length`, which captures the average linear distance between dependents and their syntactic head within a sentence. In the *Swapped* variant we simply exchanged the feature values between sentences, thus a sentence of the Standard subset that originally showed an `average link length` of, e.g., 2.86 could be changed to 8.83, a value originally associated with a different sentence. Note, in fact, that both are real values extracted from our dataset with respect to the considered feature, they have been simply randomly reassigned to a different

Table 3. Average differences between the values of linguistic features in the *Gold* dataset and each *Control* datasets for each of the 7 macro-groups.

	Random			Swapped		
Group	**Random**	**Bins**	**Lengths**	**Swapped**	**Bins**	**Lengths**
Order	0.48	0.48	0.48	0.41	0.40	0.40
POS	0.40	0.31	0.25	0.12	0.12	0.12
Subord	0.43	0.41	0.41	0.38	0.35	0.35
SyntacticDep	0.40	0.31	0.25	0.15	0.13	0.12
TreeStructure	0.36	0.28	0.25	0.20	0.18	0.18
VerbInflection	0.47	0.47	0.47	0.44	0.43	0.44
VerbPredicate	0.42	0.41	0.40	0.26	0.25	0.25
Average	0.42	0.38	0.36	0.28	0.27	0.26

Table 4. Average differences between the values of linguistic features in the *Gold* dataset and each *Control* datasets for each of the 7 macro-groups considering only the Shortest and Longest subsets.

	Shortest		Longest	
Group	**Random**	**Swapped**	**Random**	**Swapped**
Order	0.50	0.32	0.46	0.35
POS	0.43	0.13	0.38	0.13
Subord	0.49	0.20	0.39	0.28
SyntacticDep	0.44	0.13	0.37	0.15
TreeStructure	0.37	0.22	0.37	0.14
VerbInflection	0.50	0.34	0.44	0.42
VerbPredicate	0.46	0.24	0.39	0.21
Average	0.46	0.23	0.40	0.24

sentence. When building the *Random* variant, the whole sentences here considered have been associated with a feature value randomly generated between 1.33 and 9.78, which are the reported minimum and maximum `average link length` values in the dataset (Standard subset).

Since the value of many considered features is highly influenced by the length of the sentence, we defined two additional alteration strategies to be combined with the main ones that account for such a property. In the first sub-type, *Bins*, we grouped sentences falling into the same predefined range of sentence lengths (i.e., 10-15, 15-20, 20-25 and 25-30 tokens). In a second sub-type, *Lengths*, we created groups of sentences having exactly the same length. Note that we applied these strategies only to sentences from the Standard subset since the other two subsets do not present a considerable number of sentences for a given length.

Note that the different data-altering strategies are conceived to represent challenging testbeds to assess the effectiveness of our probing tasks in different scenarios. The *Swapped* control datasets are possibly the most challenging ones as the swapped feature values might be quite similar to the gold ones, thus possibly predicted with high accuracy by the probing model. Such intuition seems to be confirmed by the differences between the values of the Gold and each control dataset, obtained by averaging the differences between the gold and the altered values that each sentence had in the corresponding dataset. This holds both in the Standard (Table 3) and the Shortest and Longest subsets (Table 4). As can be noted, lower differences are reported for the Swapped control datasets, both on average and for each features group, in all subsets. Indeed, while the Random strategy tends to produce datasets where all possible values ranging between the maximum and minimum of that feature are equally distributed along sentences, the Swapped option simply shuffles gold values across sentences (namely, the mean value of a feature in the dataset does not change), producing untruthful but more plausible datasets.

3.4. Models

For all experiments, we relied on an Italian pre-trained version of the BERT model, one of the most prominent NLMs. Specifically, we used the base cased BERT developed by the MDZ Digital Library Team, available through the Huggingface's *Transformers* library [49][1]. The model was trained using the Italian Wikipedia and the OPUS corpus [50]. For obtaining the sentence-level representations for each of the 12 layers of BERT, we leveraged the activation of the first input token *[CLS]*.

The probing model is a linear Support Vector Regression model (LinearSVR). The model takes as input the above layer-wise sentence-level representations and it predicts the value of each considered feature in the Gold and Control datasets. Specifically, we trained and tested the probing model adopting a cross-validation process on each dataset individually. To this aim, we split each dataset into five portions containing the same amount of randomly selected sentences; then, we iteratively trained the probing model on four portions and used the remaining fifth as a test set. This way, the model is trained using a representative sample of the dataset at each iteration.

As an evaluation metric, we used the Spearman correlation coefficient between the values of the linguistic features in gold and control datasets and their values when predicted by the probing model using BERT's sentence-level representations as input. In the remainder of the paper, we refer to the evaluation metric as *probing score*.

Since previous work already showed the ability of pre-trained NLMs to outperform simple baselines (e.g. linear model trained using only sentence length as input feature) in the resolution of probing tasks [51], in this current paper we did not perform a direct comparison with a baseline. Nevertheless, since the focus of this work is on assessing the sensitivity of BERT to distorted feature values, control datasets can be viewed as a baseline themselves.

4. Results

Our first analysis was devoted to assessing BERT's abilities in the prediction of the authentic values of the Gold dataset. Such results represent the reference performance against which we compared the performance obtained on the diverse control datasets we built. To better appreciate the impact of sentence length as a possible confounder of the probing approach we devised, we kept separated the discussion of the results obtained on the Standard subset from the outcomes of the probing tasks performed on the Shortest and Longest subsets.

4.1. Probing on the Standard subset

As a first analysis, we probed BERT's linguistic competence with respect to the 7 groups of probing features. Figure 2 shows how the model's abilities to predict the considered linguistic phenomena in the Gold dataset change across layers. As can be noted, regardless of the group, BERT tends to lose knowledge as far as the output layer is approaching. As suggested by Liu *et al.* [31], this could be due to the fact that the representations that are better suited for language modeling are also those that exhibit worse probing task performance, indicating that Transformer layers trade-off between encoding general and probed features. However, in line with what was observed by Miaschi *et al.* [22,38] for the Italian and English languages respectively, each group of features has different behavior. Namely, the distributions of Parts-Of-Speech (*POS*) and dependency relations (*SyntacticDep*) are the best-encoded types of information, especially in the first layers, then decrease constantly. On the contrary, more complex linguistic knowledge about the order of subjects and objects with respect to the verb (*Order*) is acquired only in the middle layers. Notably, the model shows very scarce competencies about the number of dependents of a verbal head (*VerbPredicate*) which are quite constant across layers.

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

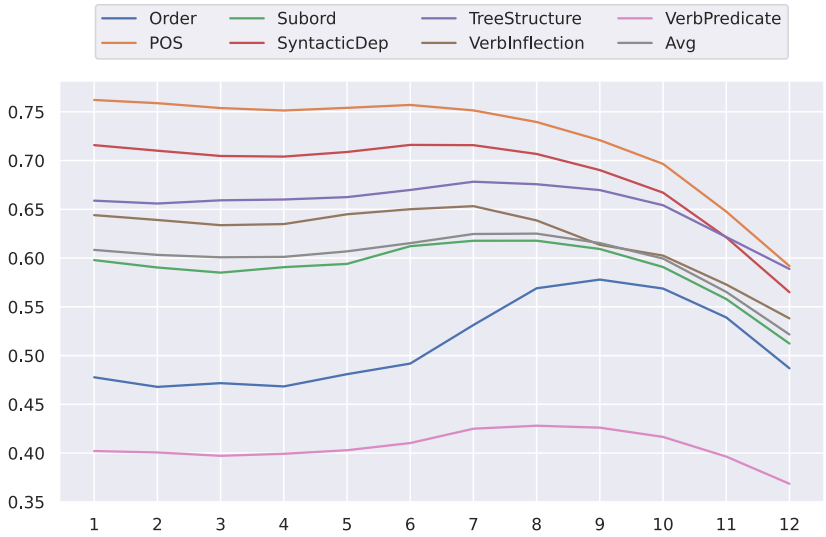

Figure 2. Layer-wise probing scores (Spearman correlations) obtained when predicting *Gold* feature values of the Standard subset according to the 7 macro-groups of linguistic features. Average results (*Avg*) are also reported.

To further investigate these trends across layers, for each feature we computed the slopes of a linear regression line between BERT layers and the values of the probing scores in the last and first layers. The *Gold* column of Figure 3 reports the slopes for the 7 groups of features and for the total amount of 77 features (line *Avg*). As it can be noted, all the slope values are negative thus indicating that the learning curve decreases across layers. The only exception is represented by the trend of the features of the *Order* group which has a positive value. It follows from the quite unique trend observed in Figure 2: the knowledge about this type of linguistic phenomenon, albeit very low, starts increasing in the middle layers, and decreases in the last ones even though it remains higher with respect to the first ones. The features that BERT tends to know quite constantly across layers are those belonging to *VerbPredicate* group. Accordingly, the slope value is the lowest one (-0.017).

Figure 3 also allows a first comparison between the performances of the probing model tested on the Gold and control datasets. The majority of negative slope values reported here show that BERT's knowledge generally tends to decrease across layers also when tested against the different typologies of control datasets[2]. Few exceptions are unevenly scattered across layers and groups of features and they are not worth discussing. However, the most striking result emerging from Figure 3 is that the slopes are quite flat both on average and considering specific features. Contrary to what was seen for the Gold dataset, we observe very small differences between the probing scores achieved using the representations extracted from the last and first layers, indicating that the knowledge about linguistic features on all control datasets is stable across layers. This result seems to suggest that altering the values of the gold features has a generic impact on BERT's linguistic knowledge.

The extent of such an impact is clear by inspecting Figure 4 which reports the gaps between the probing scores obtained when predicting gold and altered linguistic features. Here we focused on the scores achieved in the output layer since we previously observed very small changes in probing performance across layers. Specifically, the gap was computed as the difference between the probing score obtained at layer 12 on the Gold dataset and on each control dataset. Note that in order to weigh the impact of the altered feature values with respect to BERT's competence about that linguistic phenomenon, we divided the computed difference by the probing score obtained for each feature at layer 12 in the

[2] Refer to A1 for the layer-wise probing scores obtained on each control dataset.

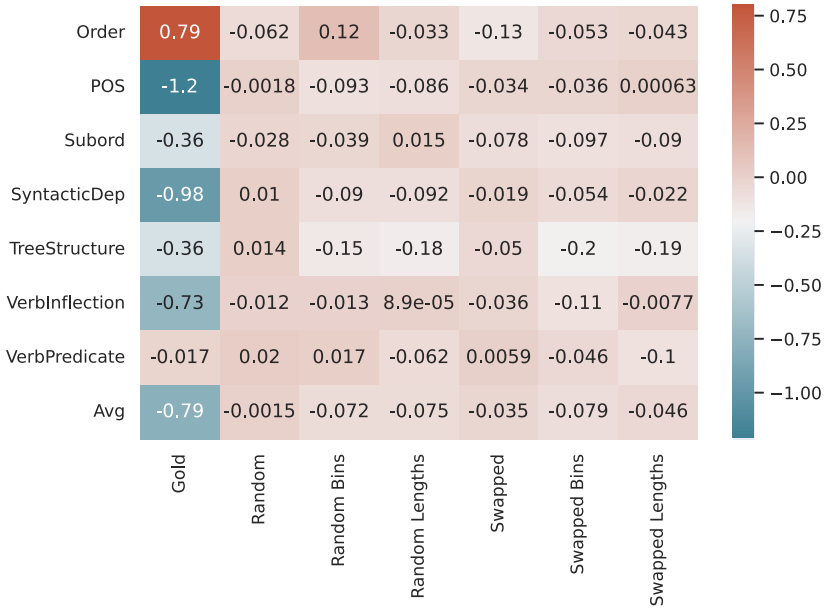

Figure 3. Slopes of the regression lines across the 12 layers for the probing scores obtained with the *Gold* and the corresponding *Control* datasets. Scores are multiplied by 100.

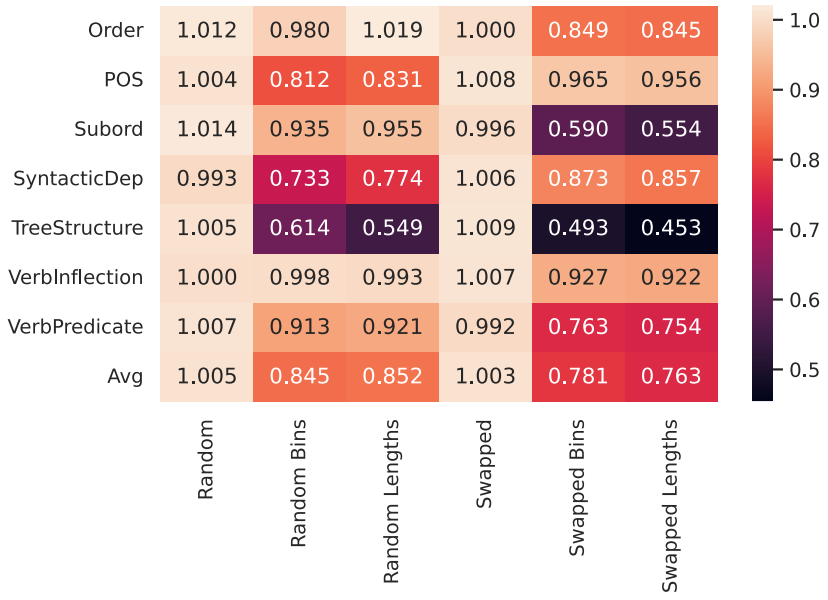

Figure 4. Differences between the probing scores obtained with the *Gold* dataset and each *Control* dataset using BERT representations extracted from the output (12) layer.

Gold dataset [3]. Differences higher than 1 are obtained when the probing scores achieved on the control dataset are lower than 0.

The positive value of differences visualized in the heatmap shows that on average (*Avg* row), and for all groups of features, the highest probing scores are obtained on the Gold dataset even with some differences across the typologies of features and control datasets. The greatest differences are obtained for the *Random* and *Swapped* datasets without any constraints about the length of sentences. This seems to suggest that the probing model is able to recognize that the feature values contained in the two main types of control datasets

[3] The formula adopted for every single feature is the following one: (probing score at layer 12 in the Gold dataset - probing score at layer 12 in the control dataset) / probing score at layer 12 in the Gold dataset

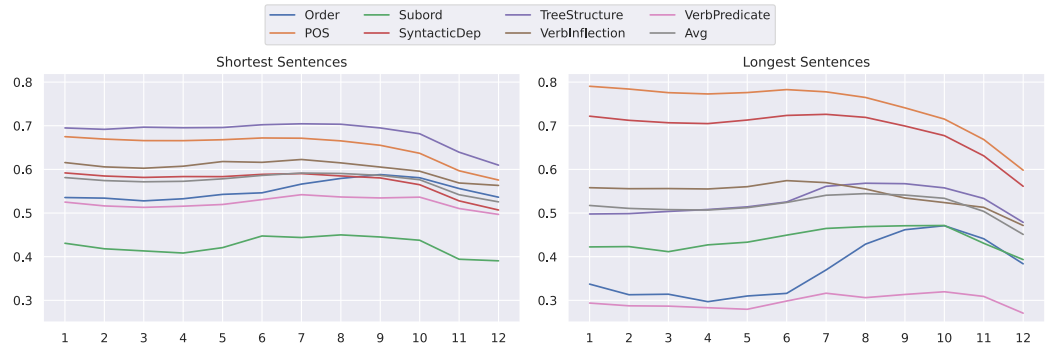

Figure 5. Layer-wise probing scores (Spearman correlations) obtained when predicting *Gold* feature values according to the 7 macro-groups of linguistic features for the *Shortest* and *Longest* subsets. Average results (*Avg*) are also reported.

have been altered, even when they are not fully random but plausible, i.e. in the *Swapped* datasets. As a consequence, we can hypothesize that the probing model is relying on some implicit linguistic knowledge when it predicts the authentic feature values, rather than learning some regularities possibly found in the dataset.

However, if we take a closer look at the gaps between the Gold and the altered datasets when we constrain the length of the sentences, we can observe that on average (*Avg* row) the differences with respect to the prediction of the authentic feature values are generally lower. More specifically, the *Swapped Bins* (diff=0.781) and *Lengths* (diff=0.763) datasets result to be more challenging for our probing approach than the corresponding *Random* ones, against which we obtained higher differences equal to 0.845 and 0.852, respectively. Namely, since the feature values artificially created simply by shuffling gold ones across sentences constrained by sentence length are more similar to the gold values, as shown in Table 3, the swapped values result to be more confounding for the probing model. In fact, they are predicted with higher accuracy than randomly altered values.

In addition, stronger differences across groups of features emerge from this analysis. BERT's generalization abilities of features referring to the local and global syntactic structure of the sentence (*TreeStructure*) seem the most similar to gold ones based on the relatively small gap between predictions. Note that these sentence properties are the most sensitive to the sentence length, which BERT encodes with very high accuracy [52]. This may suggest that in the resolution of these tasks the probing model is possibly relying on some regularities related to sentence length. The same holds for features related to *Subordination*, which are similarly highly correlated to sentence length. On the contrary, in both *Swapped* and *Random* control datasets, the probing model performances diverge with respect to the prediction of the pre- or post-verbal order of subject and object (*Order*) in a sentence, and in particular of the verbal morphology features (*VerbInflection*), as shown by their smaller gaps.

4.2. Probing on the Shortest and Longest subsets

In this section, we take a closer look at how BERT performs when tested against the *Shortest* and *Longest* subsets of IUDT sentences, which, as described in Section 3.1, gather all sentences having a length up to 9 tokens and between 31 and 100 tokens, respectively. As in the previous section, we start by reporting the layer-wise probing scores obtained by the model when predicting the gold values of the linguistic features extracted from sentences belonging to these two subsets. This is shown in Figure 5, where we can see how BERT's implicit knowledge changes across layers and groups of linguistic phenomena. A first observation that we can draw from the figure is that the subset of long sentences exhibits a higher variation across layers, and that this trend is more similar to the one observed for the *Standard* subset (see Figure 2). This holds especially for some groups of phenomena, such as the distributions of Parts-Of-Speech (*POS*) and of dependency

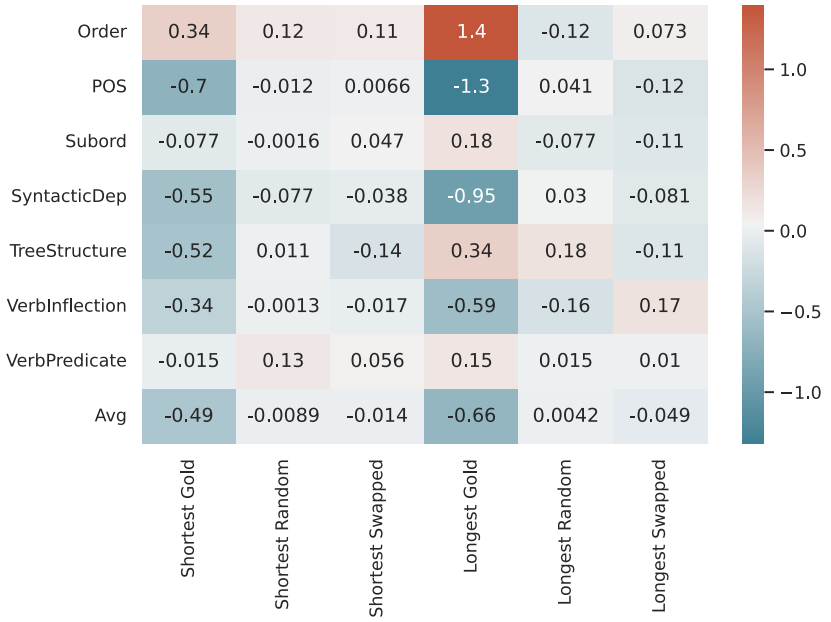

Figure 6. Slopes of the regression lines across the 12 layers for the probing scores obtained with the *Gold* and the corresponding *Control* datasets for the Shortest and Longest subsets. Scores are multiplied by 100.

relations (*SyntacticDep*), for which BERT's predictions are very similar to the gold value, especially in the first layers, whereas this specific knowledge tends to decrease as the output layer is approached. A further similarity can be observed with respect to the worst encoded features, which are represented by sentence properties related to the complexity of verbal predicates (*VerbPredicate*) and, although to a lesser extent, to syntactic ordering (*Order*). Note that the latter group, as already observed for the *Standard* subset, is better encoded in the middle layers rather than in the first ones.

On the contrary, BERT's linguistic knowledge tested on the subset of short sentences is on average more stable with few variations across layers and across the diverse groups of linguistic phenomena. In addition, we can observe that BERT competencies are differently ranked with respect to the ranking obtained in the *Standard* and *Longest* subsets. In fact, the features that the language model masters with the highest accuracy are those modeling the syntactic structure of the sentence (*TreeStructure*) with a layer-wise average probing score equal to 0.68. Note that such a score is higher than the accuracy achieved in the *Longest* (0.53) and *Standard* (0.65) subsets. This result may be a consequence of the fact that the values of the features belonging to this group are highly sensitive to sentence length and short sentences are typically characterized by quite flat, and simple, syntactic trees (as shown in Table 2). Since, as we mentioned, sentence length is a feature that BERT masters very well, BERT may rely on the knowledge of this shallow feature as a proxy to predict more complex features related to the structure of the syntactic tree. Possibly related to the same reason, it results that BERT masters the order of subject and object (*Order*), and the number of dependents of verbal heads (*VerbPredicate*), much better in short than in long sentences, with accuracies even higher than the ones achieved on the Standard subset[4]. Differently from the other two datasets, the worst prediction is achieved by the features modeling the subordination (*Subord*), even if with probing scores very similar to the ones of the *Longest* subset.

[4] Layer-wise average probing scores of the *Order* group are 0.55 in the *Shortest* subset, 0.37 in the *Longest*, and 0.51 in the *Standard* one. The scores achieved for the *VerbPredicate* group are 0.52, 0.29 and 0.40 in the three datasets respectively.

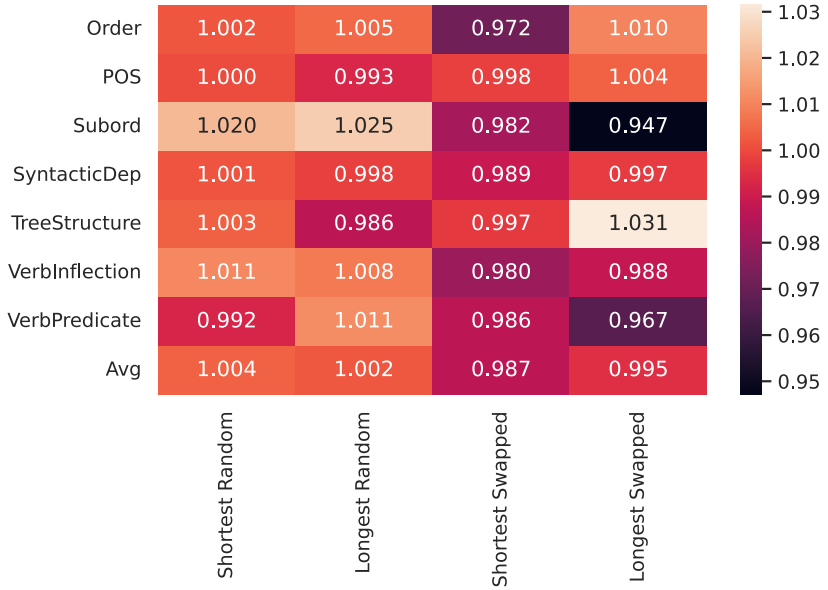

Figure 7. Differences between the probing scores obtained for the *Gold* dataset and each *Control* dataset for the *Longest* and *Shortest* subsets. Differences are computed using BERT representations extracted from the output (12) layer.

Despite these differences, Figure 6 shows that BERT's knowledge tends to change very little across layers with respect to what was observed for the sentences in the *Standard* subset[5]. As previously noted, the average flattest slopes are obtained considering the Shortest subset (*Avg*=-0.49), while more variations can be seen for the Longest one. It is also worth highlighting that in the latter case we have several groups of features with positive slope values. This is the case not only of the features belonging to the *Order* group, which have the same trend in the *Standard* and *Shortest* subsets but also of the features modeling the subordination, the syntactic structure of a sentence and the verbal arity.

In addition, the figure allows a first analysis of the impact of the corresponding control datasets on the probing model performance. Specifically, we can see that the altered feature values are predicted quite similarly across layers, while the prediction of the gold values undergoes more variations. This trend is similar to the one reported for the Standard subsets and it suggests that also in less standard sentences the probing model is sensitive to the distortion of feature values. Further evidence in this direction can be acquired by inspecting Figure 7 which reports the gap between the probing model accuracy on the *Gold* and *Control* datasets for the two subsets. As noted in the previous section, the positive values show that the gold values of the features are predicted with higher accuracies than the altered ones[6]. As in the case of the *Standard* subset (see Figure 4), very few variations across the groups of features emerge, thus showing that the probing model is scarcely confused by the distortion of feature values regardless of the linguistic phenomenon tested. We noticed for example that even if BERT's knowledge concerning the use of subordination is lower both in the *Shortest* and *Longest* subsets than in the *Standard* one. However, the gap between the probing scores obtained for the corresponding *Gold* and *Control* datasets is similarly high in the three subsets. However, differently from what we observed for the *Standard* subset, the *Swapped* control datasets are slightly more challenging than the *Random* ones. In fact, the differences are on average (*Avg* row) lower, especially when the probing model is tested against the control datasets of the short sentences.

[5] Refer to A2 for the layer-wise probing scores obtained on each control dataset.

[6] Also in this case, the differences are weighted based on the probing scores obtained by each feature on the gold *Shortest* and *Longest* subsets.

5. Conclusion

In spite of the wide amount of studies that have relied on the diagnostic probing paradigm to assess the linguistic knowledge implicitly encoded by NLM's representations, the validity of this method is still questionable from different perspectives. Our study has presented a novel contribution to this debate by focusing specifically on one of the still open questions, that is the effectiveness of probes to reflect the linguistic properties encoded in a representation. To this aim, we analyzed the performance of a probing model trained with layer-wise sentence-level BERT's representations to predict the value of a large set of linguistic features derived from the Italian Universal Dependency Treebank (IUDT) and from a suite of control datasets specifically created to alter the original value of the examined features.

As a general remark, we observed that the probing model has always a better performance when tested against the IUDT datasets than against the corresponding control datasets. Namely, the gold values of the considered set of linguistic features are predicted with higher accuracy than the artificially altered ones, thus showing that the probing model is sensitive to the distortion of feature values and it does not simply learn the regularities of the probing task. Such a result corroborates the reliability of the probing task as an interpretability approach to assess the level of linguistic knowledge implicitly encoded in BERT's sentence-level representations.

However, our experiments also highlighted that sentence length is a relevant confounding factor that may bias the real estimate of BERT's linguistic knowledge. In fact, when we focused on sentences of the same length (or same ranges of lengths) taken from the *Standard* IUDT subset, we observed that the probing model is less sensitive to the artificially generated values of features, especially when these values were obtained by shuffling the original values across sentences of the same length (or ranges of lengths). This suggests that, when the length is controlled, an alteration strategy that assigns incorrect but still plausible values is more challenging for the probing model than one that simply generates random values. Such a general trend concerns in particular the groups of linguistic phenomena that are more influenced by the length of the sentence. This is the case for example of the features modeling local and global characteristics of the syntactic structure of a sentence (i.e. the *TreeStructure* group) which tend to have quite homogeneous values within sentences of the same length. Accordingly, the output space of the probing model for these features is smaller than in the whole dataset, thus making them more easily predictable without relying on authentic linguistic competence. Despite this trend being particularly visible when we consider the output layer, we showed that the probing model is sensitive to the altered values also across the twelve BERT's layers. Quite interestingly, contrary to what was observed for the Gold dataset, the learning curve of the model tested on the control datasets decreases quite slowly across layers, with no significant variations across the typology of linguistic phenomena.

The main outcomes obtained for the group of sentences with a standard length in Italian were also confirmed by the experiments conducted on the subsets of sentences having less standard lengths. Although we highlighted that BERT masters specific linguistic phenomena with different accuracy in *Shortest*, *Longest* and *Standard* subsets, we showed that the probing model is similarly scarcely confused in the three subsets regardless of the linguistic aspect considered. This seems to suggest that BERT's representations extracted from less standard sentences implicitly encode the linguistic knowledge of therein phenomena.

The present study can be extended from diverse perspectives. In the future, the effectiveness of the diagnostic probing approach can be evaluated considering other languages, possibly belonging to different language families and thus characterized by different feature values. Indeed, it could be worth exploring whether confounding factors affecting the performance of probing models are shared among languages or vary depending on their family. In this respect, we can either reuse the same set of linguistic features or focus on subsets of phenomena of particular interest for a typological study. In fact, the approach

adopted to select the set of linguistic features is multilingual being based on the Universal Dependencies formalism. In addition, as neural models continue to improve, a further possible direction of research may consist in assessing the effectiveness of the probing approach to test the linguistic knowledge encoded in models with different architectures.

Appendix A

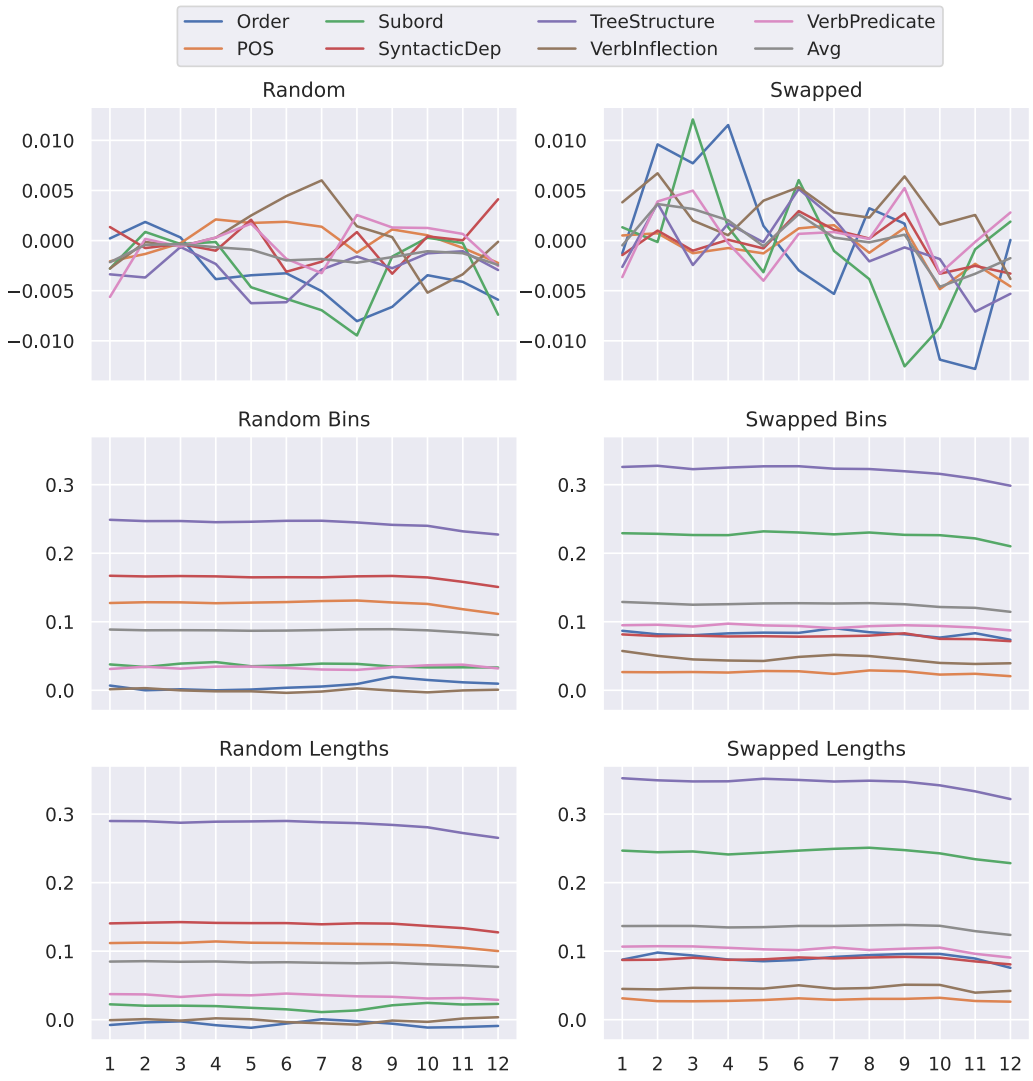

Figure A1. Layer-wise probing scores (Spearman correlations) obtained when predicting *Control* feature values of the Standard subset according to the 7 macro-groups of linguistic features. Average results (*Avg*) are also reported.

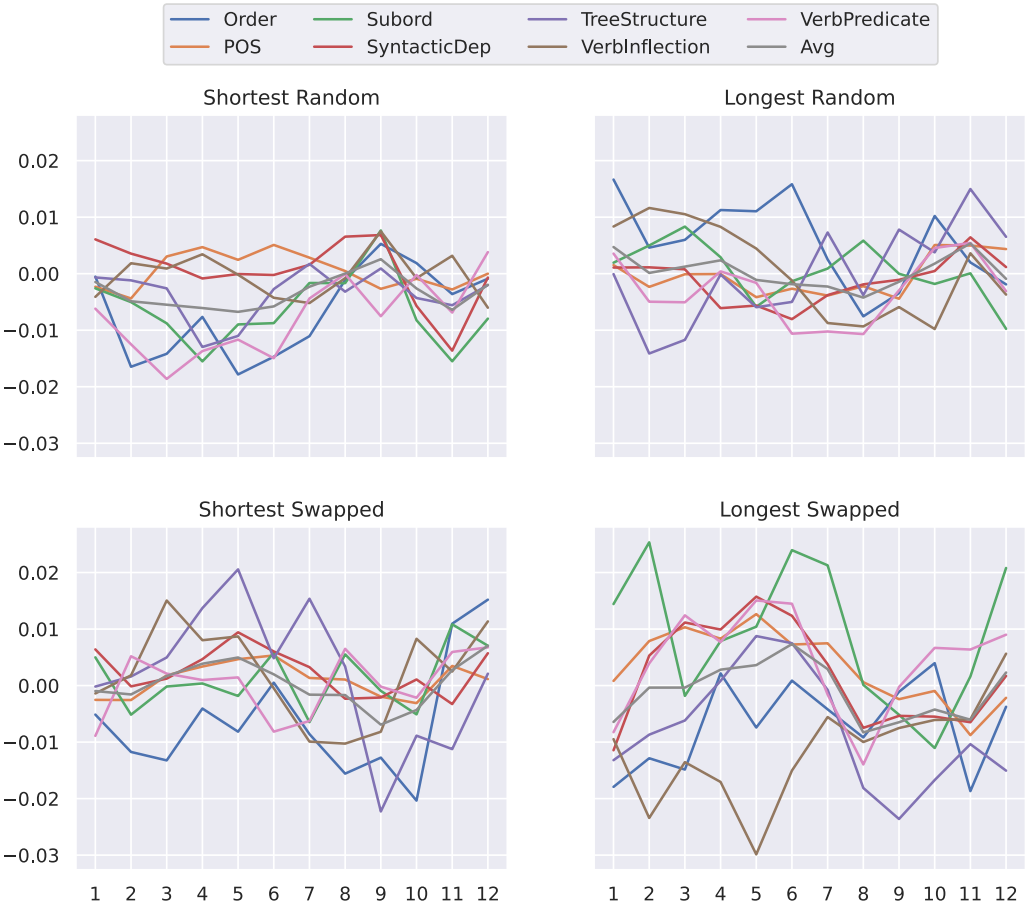

Figure A2. Layer-wise probing scores (Spearman correlations) obtained when predicting *Control* feature values according to the 7 macro-groups of linguistic features for the *Shortest* and *Longest* subsets. Average results (*Avg*) are also reported.

References

1. Miaschi, A.; Alzetta, C.; Brunato, D.; Dell'Orletta, F.; Venturi, G. Probing Tasks Under Pressure. In Proceedings of the Proceedings of the Eighth Italian Conference on Computational Linguistics (CLiC-it 2021); Fersini, E.; Passarotti, M.; Patti, V., Eds.; CEUR Workshop Proceedings (CEUR-WS.org): Milan, 2022.
2. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.
3. Wang, A.; Pruksachatkun, Y.; Nangia, N.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; Bowman, S. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems* **2019**, *32*.
4. Yang, W.; Xie, Y.; Lin, A.; Li, X.; Tan, L.; Xiong, K.; Li, M.; Lin, J. End-to-End Open-Domain Question Answering with BERTserini. In Proceedings of the Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations); Association for Computational Linguistics: Minneapolis, Minnesota, 2019; pp. 72–77. https://doi.org/10.18653/v1/N19-4013.
5. Naseem, U.; Razzak, I.; Musial, K.; Imran, M. Transformer based deep intelligent contextual embedding for twitter sentiment analysis. *Future Generation Computer Systems* **2020**, *113*, 58–69.
6. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692* **2020**.
7. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In Proceedings of the International Conference on Learning Representations, 2020.
8. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners **2019**.
9. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 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); Association for Computational Linguistics: Minneapolis, Minnesota, 2019; pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423.

10. Rogers, A.; Kovaleva, O.; Rumshisky, A. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics* **2020**, *8*, 842–866. https://doi.org/10.1162/tacl_a_00349.
11. Belinkov, Y.; Màrquez, L.; Sajjad, H.; Durrani, N.; Dalvi, F.; Glass, J. Evaluating Layers of Representation in Neural Machine Translation on Part-of-Speech and Semantic Tagging Tasks. In Proceedings of the Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2017, pp. 1–10.
12. Ettinger, A. What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics* **2020**, *8*, 34–48, [https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00298/1923116/tacl_a_0 https://doi.org/10.1162/tacl_a_00298.
13. Morger, F.; Brandl, S.; Beinborn, L.; Hollenstein, N. A Cross-lingual Comparison of Human and Model Relative Word Importance. In Proceedings of the Proceedings of the 2022 CLASP Conference on (Dis)embodiment; Association for Computational Linguistics: Gothenburg, Sweden, 2022; pp. 11–23.
14. Clark, K.; Khandelwal, U.; Levy, O.; Manning, C.D. What Does BERT Look at? An Analysis of BERT's Attention. In Proceedings of the Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP; Association for Computational Linguistics: Florence, Italy, 2019; pp. 276–286. https://doi.org/10.18653/v1/W19-4828.
15. Goldberg, Y. Assessing BERT's syntactic abilities. *arXiv preprint arXiv:1901.05287* **2019**.
16. Ramnath, S.; Nema, P.; Sahni, D.; Khapra, M.M. Towards Interpreting BERT for Reading Comprehension Based QA. In Proceedings of the Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); Association for Computational Linguistics: Online, 2020; pp. 3236–3242. https://doi.org/10.18653/v1/2020.emnlp-main.261.
17. Conneau, A.; Kruszewski, G.; Lample, G.; Barrault, L.; Baroni, M. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In Proceedings of the Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Association for Computational Linguistics: Melbourne, Australia, 2018; pp. 2126–2136. https://doi.org/10.18653/v1/P18-1198.
18. Belinkov, Y. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics* **2021**, pp. 1–12, [https://direct.mit.edu/coli/article-pdf/doi/10.1162/coli_a_00422/1965357/coli_a_00422.pdf]. https://doi.org/10.1162/coli_a_00422.
19. Zeman, D.; Nivre, J.; Abrams, M.; Aepli, N.; Agic, Ž.; Ahrenberg, L.; et al. Universal dependencies 2.5. *LINDAT/CLARIAHCZ digital library at the Institute of Formal and Applied Linguistics (UFAL), Faculty of Mathematics and Physics, Charles University. url: http://hdl. handle. net/11234/1-3226* **2020**.
20. Belinkov, Y.; Glass, J. Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics* **2019**, *7*, 49–72. https://doi.org/10.1162/tacl_a_00254.
21. Hewitt, J.; Liang, P. Designing and Interpreting Probes with Control Tasks. In Proceedings of the Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 2733–2743.
22. Miaschi, A.; Brunato, D.; Dell'Orletta, F.; Venturi, G. Linguistic Profiling of a Neural Language Model. In Proceedings of the Proceedings of the 28th International Conference on Computational Linguistics; International Committee on Computational Linguistics: Barcelona, Spain (Online), 2020; pp. 745–756. https://doi.org/10.18653/v1/2020.coling-main.65.
23. Raganato, A.; Tiedemann, J. An analysis of encoder representations in transformer-based machine translation. In Proceedings of the Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Association for Computational Linguistics, 2018.
24. Htut, P.M.; Phang, J.; Bordia, S.; Bowman, S.R. Do attention heads in BERT track syntactic dependencies? *arXiv preprint arXiv:1911.12246* **2019**.
25. Kovaleva, O.; Romanov, A.; Rogers, A.; Rumshisky, A. Revealing the Dark Secrets of BERT. In Proceedings of the Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); Association for Computational Linguistics: Hong Kong, China, 2019; pp. 4365–4374. https://doi.org/10.18653/v1/D19-1445.
26. Saphra, N.; Lopez, A. Understanding Learning Dynamics Of Language Models with SVCCA. In Proceedings of the 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), 2019, pp. 3257–3267.
27. Blevins, T.; Levy, O.; Zettlemoyer, L. Deep RNNs Encode Soft Hierarchical Syntax. In Proceedings of the Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2018, pp. 14–19.
28. Tenney, I.; Xia, P.; Chen, B.; Wang, A.; Poliak, A.; McCoy, R.T.; Kim, N.; Van Durme, B.; Bowman, S.R.; Das, D.; et al. What do you learn from context? probing for sentence structure in contextualized word representations. *Proceedings of the 7th International Conference on Learning Representations (ICLR 2019)* **2019**.
29. Hewitt, J.; Manning, C.D. A structural probe for finding syntax in word representations. In Proceedings of the 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), 2019, pp. 4129–4138.
30. Tenney, I.; Das, D.; Pavlick, E. BERT Rediscovers the Classical NLP Pipeline. In Proceedings of the Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; Association for Computational Linguistics: Florence, Italy, 2019; pp. 4593–4601. https://doi.org/10.18653/v1/P19-1452.

31. Liu, N.F.; Gardner, M.; Belinkov, Y.; Peters, M.E.; Smith, N.A. Linguistic Knowledge and Transferability of Contextual Representations. In Proceedings of the 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); Association for Computational Linguistics: Minneapolis, Minnesota, 2019; pp. 1073–1094. https://doi.org/10.18653/v1/N19-1112.
32. Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep Contextualized Word Representations. In Proceedings of the Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers); Association for Computational Linguistics: New Orleans, Louisiana, 2018; pp. 2227–2237. https://doi.org/10.18653/v1/N18-1202.
33. Hall Maudslay, R.; Valvoda, J.; Pimentel, T.; Williams, A.; Cotterell, R. A Tale of a Probe and a Parser. In Proceedings of the Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; Association for Computational Linguistics: Online, 2020; pp. 7389–7395. https://doi.org/10.18653/v1/2020.acl-main.659.
34. Pimentel, T.; Valvoda, J.; Maudslay, R.H.; Zmigrod, R.; Williams, A.; Cotterell, R. Information-Theoretic Probing for Linguistic Structure. In Proceedings of the Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4609–4622.
35. Voita, E.; Titov, I. Information-Theoretic Probing with Minimum Description Length. In Proceedings of the Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 183–196.
36. Ravichander, A.; Belinkov, Y.; Hovy, E. Probing the Probing Paradigm: Does Probing Accuracy Entail Task Relevance? In Proceedings of the Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume; Association for Computational Linguistics: Online, 2021; pp. 3363–3377.
37. de Vries, W.; van Cranenburgh, A.; Nissim, M. What's so special about BERT's layers? A closer look at the NLP pipeline in monolingual and multilingual models. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020; Association for Computational Linguistics: Online, 2020; pp. 4339–4350. https://doi.org/10.18653/v1/2020.findings-emnlp.389.
38. Miaschi, A.; Sarti, G.; Brunato, D.; Dell'Orletta, F.; Venturi, G. Italian Transformers Under the Linguistic Lens. In Proceedings of the Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020); Monti, J.; Dell'Orletta, F.; Tamburini, F., Eds.; CEUR Workshop Proceedings (CEUR-WS.org): Online, 2021.
39. Guarasci, R.; Silvestri, S.; De Pietro, G.; Fujita, H.; Esposito, M. Assessing BERT's ability to learn Italian syntax: a study on null-subject and agreement phenomena. *Journal of Ambient Intelligence and Humanized Computing* **2021**, pp. 1–15.
40. Sanguinetti, M.; Bosco, C. Converting the parallel treebank ParTUT in Universal Stanford Dependencies. *Converting the parallel treebank ParTUT in Universal Stanford Dependencies* **2014**, pp. 316–321.
41. Delmonte, R.; Bristot, A.; Tonelli, S. VIT-Venice Italian Treebank: Syntactic and Quantitative Features. In Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories. Northern European Association for Language Technol, 2007, Vol. 1, pp. 43–54.
42. Bosco, C.; Simonetta, M.; Maria, S. Converting italian treebanks: Towards an italian stanford dependency treebank. In Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse. The Association for Computational Linguistics, 2013, pp. 61–69.
43. Zeman, D.; Popel, M.; Straka, M.; Hajič, J.; Nivre, J.; Ginter, F.; Luotolahti, J.; Pyysalo, S.; Petrov, S.; Potthast, M.; et al. CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In Proceedings of the Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies; Association for Computational Linguistics: Vancouver, Canada, 2017; pp. 1–19. https://doi.org/10.18653/v1/K17-3001.
44. Sanguinetti, M.; Bosco, C.; Lavelli, A.; Mazzei, A.; Antonelli, O.; Tamburini, F. PoSTWITA-UD: an Italian Twitter Treebank in universal dependencies. In Proceedings of the Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018.
45. Cignarella, A.T.; Bosco, C.; Patti, V.; Lai, M. TWITTIRÒ: an Italian Twitter Corpus with a Multi-layered Annotation for Irony. *IJCoL. Italian Journal of Computational Linguistics* **2018**, *4*, 25–43.
46. Miaschi, A.; Brunato, D.; Dell'Orletta, F.; Venturi, G. Linguistic Profiling of a Neural Language Model. In Proceedings of the Proceedings of the 28th International Conference on Computational Linguistics; International Committee on Computational Linguistics: Barcelona, Spain (Online), 2020; pp. 745–756. https://doi.org/10.18653/v1/2020.coling-main.65.
47. Brunato, D.; Cimino, A.; Dell'Orletta, F.; Venturi, G.; Montemagni, S. Profiling-UD: a Tool for Linguistic Profiling of Texts. In Proceedings of the Proceedings of The 12th Language Resources and Evaluation Conference; European Language Resources Association: Marseille, France, 2020; pp. 7147–7153.
48. Nivre, J. Towards a universal grammar for natural language processing. In Proceedings of the Proceedings of The 16th Annual Conference on Intelligent Text Processing and Computational Linguistics (CICLing). Springer, 2015, pp. 3–16.
49. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations; Association for Computational Linguistics: Online, 2020; pp. 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6.

50. Tiedemann, J.; Nygaard, L. The OPUS Corpus - Parallel and Free: http://logos.uio.no/opus. In Proceedings of the Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04); European Language Resources Association (ELRA): Lisbon, Portugal, 2004.
51. Miaschi, A.; Sarti, G.; Brunato, D.; Dell'Orletta, F.; Venturi, G. Probing Linguistic Knowledge in Italian Neural Language Models across Language Varieties. *IJCoL. Italian Journal of Computational Linguistics* **2022**, *8*.
52. Jawahar, G.; Sagot, B.; Seddah, D. What Does BERT Learn about the Structure of Language? In Proceedings of the Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; Association for Computational Linguistics: Florence, Italy, 2019; pp. 3651–3657. https://doi.org/10.18653/v1/P19-1356.