Probing Tasks Under Pressure

Alessio Miaschi[1,2], **Chiara Alzetta**[1], **Dominique Brunato**[1],
Felice Dell'Orletta[1], **Giulia Venturi**[1]
[1]Istituto di Linguistica Computazionale "Antonio Zampolli", Pisa
ItaliaNLP Lab – *www.italianlp.it*
[2]Department of Computer Science, University of Pisa
alessio.miaschi@phd.unipi.it, name.surname@ilc.cnr.it

Abstract

Probing tasks are frequently used to evaluate whether the representations of Neural Language Models (NLMs) encode linguistic information. However, it is still questioned if probing classification tasks really enable such investigation or they simply hint for surface patterns in the data. We present a method to investigate this question by comparing the accuracies of a set of probing tasks on gold and automatically generated control datasets. Our results suggest that probing tasks can be used as reliable diagnostic methods to investigate the linguistic information encoded in NLMs representations.

1 Introduction

In recent years we saw the raise of a consistent body of work dealing with the use of probing tasks to test the linguistic competence learned by Neural Language Models (NLMs) (Conneau et al., 2018; Warstadt et al., 2019; Hewitt and Liang, 2019; Miaschi et al., 2020). The idea behind the probing paradigm is actually quite simple: using a diagnostic classifier, the *probing model* or *probe*, that 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 method reported that NLMs representations do encode several properties related to morphological, syntactic and semantic information.

Despite the amount of work, there are still several open questions concerning their use (Belinkov, 2021): which probing model should we use for assessing the linguistic competence of a NLM? Are probes the most effective strategy to achieve such 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 the effectiveness of probing models. Both are still not well investigated issues, although their importance for advancing the research on the evaluation of NLMs linguistic competences has been widely recognized.

Among the first line of research, dealing 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 (Hewitt and Manning, 2019; Liu et al., 2019; Hall Maudslay et al., 2020) or complex ones (Pimentel et al., 2020; Voita and Titov, 2020) in terms of model parametrization. Specifically, Voita and Titov (2020) suggest to design alternative probes using a novel information-theoretic approach which balances the probe inner complexity with its task performance.

Concerning works facing the issue of investigating the effectiveness of the probing paradigm, Hewitt and Liang (2019) observe 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 idea, they introduced *control tasks*, a set of tasks that associate word types with random outputs that can be solved by simply learning regularities. Along the same line, Ravichander et al. (2021) 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.

The scenario defined by the latter two works is the one we deal with in this paper. Specifically, we introduce a new approach to put increasingly under pressure the effectiveness of a suite of probing tasks to test the linguistic knowledge implicitly encoded by BERT (Devlin et al., 2019), one of the most prominent NLMs. To achieve this goal, we set up a number of experiments (see Section 2) aimed at comparing the performance of a regression model trained with BERT representations to predict the values of a set of linguistic properties extracted from the Italian Universal Dependency Treebank (Zeman et al., 2020) and from a suite of *control datasets* we specifically built for the purpose of this study. We define a control dataset as a set of linguistic features whose values were automatically altered in order to be increasingly different from the values in the treebank, referred to as *gold* values. Our underlying hypothesis is that if the predictions of the increasingly altered values progressively 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 representation We will discuss the results of our experiments in light of this hypothesis in Section 3. In Section 4 we will draw the conclusions.

Note that this is one of the few studies focused on non-English NLMs. In fact, with the exception of (de Vries et al., 2020; Miaschi et al., 2021; Guarasci et al., 2021), the majority of research related to interpretability issues is focused on English or, at most, multilingual models.

Contributions To the best of our knowledge this is the first paper that (i) introduces a methodology to test the reliability of probing tasks by building control tasks at increasing level of complexity, (ii) puts under pressure the probing approach considering the Italian language.

2 Methodology

Our methodology seeks to investigate the effectiveness of probing tasks for evaluating the linguistic competences encoded in NLM representations. To this aim, we trained a probing model (described in Section 2.1) using BERT sentence representations and then tested its performance when predicting the values of a set of linguistic features (see Section 2.3) in multiple scenarios. In one scenario, the model shall predict gold values, thus corresponding to the real values of the features in the corpus. In the other scenarios, we automatically altered the feature values at different control levels each corresponding to increasing degrees of pressure for the probing model, as discussed in Section 2.4.

Our methodology will allow us to test whether the probing model really encodes linguistic competences or simply learns regularities in the task and data distributions by checking the results obtained in the different scenarios. If the predictions of the probing model will be more similar to the gold values than to the automatically altered ones, then we might assume that the information captured by the probed feature is encoded in the representations.

2.1 Model

Our model is a pre-trained Italian BERT. Specifically, we used the base cased BERT developed by the MDZ Digital Library Team, available trough the Huggingface's *Transformers* library (Wolf et al., 2020)[1]. The model was trained using Wikipedia and the OPUS corpus (Tiedemann and Nygaard, 2004). For the sentence-level representations, we leveraged the activation of the first input token *[CLS]*. The probing model is a linear Support Vector Regression model (LinearSVR).

2.2 Data

Our experiments are carried out on the Italian Universal Dependencies Treebank (IUDT), version 2.5 (Zeman et al., 2020), containing a total of 35,480 sentences. Due to the IUDT high variability in terms of sentence length[2], we focused on a sub-set of sentences with a ± 10 tokens variation with respect to the median sentence length (i.e. 20 tokens). As a result, we selected 21,991 sentences whose length ranges between 10 and 30 tokens. This way our dataset is balanced, viz., the amount of sentences with exact same length considered for the experiments is comparable. Specifically, our dataset accounts for around 1,000 sentences for each reported value of sentence length, which makes the results of our analyses reliable and comparable.

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

[2] IUDT contains sentences ranging from 1 to 308 token long.

Morphosyntactic information
Distibution of UD POS
Lexical density
Inflectional morphology
Distribution of lexical verbs and auxiliaries for inflectional categories (tense, mood, person, number)
Verbal Predicate Structure
Distribution of verbal heads and verbal roots
Average verb arity and distribution of verbs by arity
Global and Local Parsed Tree Structures
Depth of the whole syntactic tree
Average length of dependency links and of the longest link
Average length of prepositional chains and distribution by depth
Average clause length
Relative order of elements
Distribution of subjects and objects in post- and pre-verbal position
Syntactic Relations
Distribution of dependency relations
Use of Subordination
Distribution of subordinate and principal clauses
Average length of subordination chains and distribution by depth
Distribution of subordinates in post- and pre-principal clause position

Table 1: Linguistic features probed in the experiments.

2.3 Linguistic Features

The probing tasks we defined consist in predicting the value of multiple linguistic features, each corresponding to a specific property of sentence structure. The set includes 77 linguistic features and it is based on the ones described in Brunato et al. (2020) modeling 7 main aspects of the structure of a sentence, which are reported in Table 1. They 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 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 a variety of classification problems where the linguistic information plays a fundamental role. In addition, they are multilingual as they are based on the Universal Dependency formalism for sentence representation (Nivre, 2015). In fact, they have been successfully used to profile the knowledge encoded in the language representations of contextual NLMs for both the Italian (Miaschi et al., 2021) and English language (Miaschi et al., 2020).

In this study, the values of each feature acquired from IUDT represent the *gold dataset* and they have been automatically altered in order to generate additional *control datasets*.

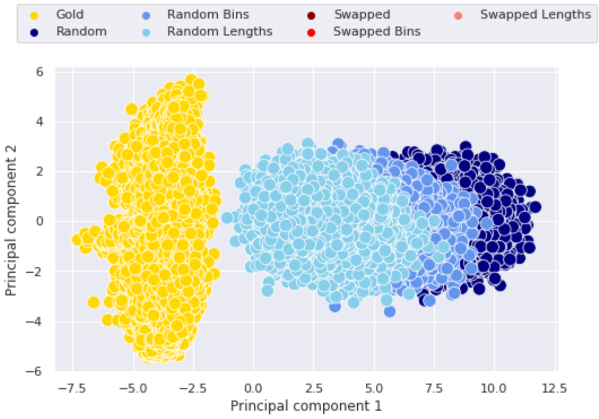

Figure 1: 2-dimensional PCA projection of the feature values in the gold and control datasets. All *Swapped* datasets overlap with the *Gold* one.

2.4 Control Datasets

We created two main types of control datasets, obtained by automatically altering gold feature values. The first main type (hereafter referred to as *Swapped*) is built by shuffling the original values of each feature across sentences; while the second type (*Random*) contains values randomly generated within the maximum and the minimum value that each feature shows in the whole gold dataset. 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 swap the feature values, thus a sentence which originally showed an `average link length` of, e.g., 2.86 could be changed to 8.83. Note that both are real values extracted from our dataset. When building the *Random* variant, all sentences considered for the study show 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, respectively associated to sentences with length 11 and 21.

Since the values of the considered features are strongly related to the length of the sentence, for each type of control dataset we built two sub-types of datasets. In a 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 included groups of sentences having exactly the same length. This motivates the choice of

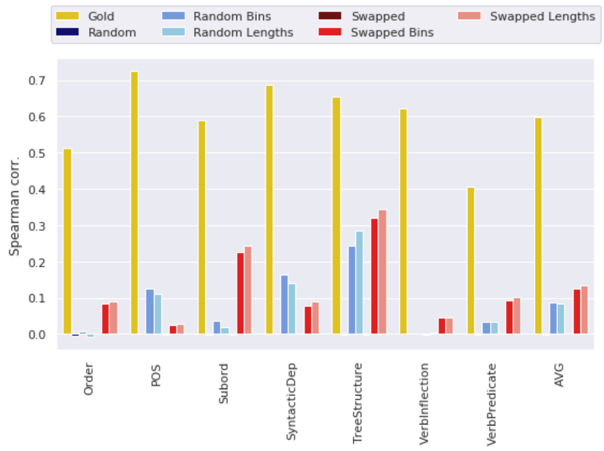

Figure 2: Average probing scores (as Spearman correlation) obtained by the LinearSVR model when predicting *gold* and *control* linguistic features. Results are reported for each feature group and on average ('AVG' column).

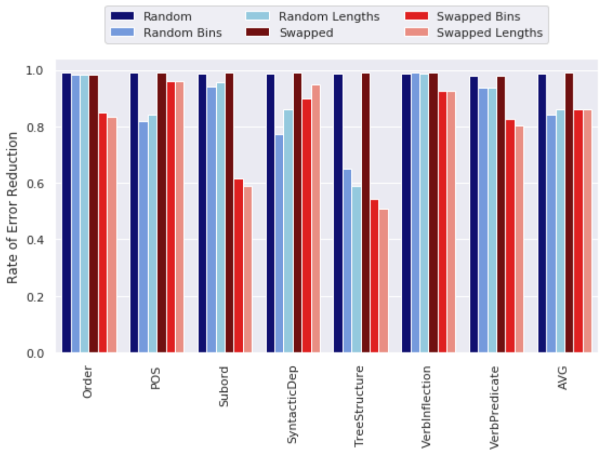

Figure 3: Error reduction rates reporting the difference between the probing scores obtained on the Gold dataset and each control dataset. Result are reported for each feature group and on average ('AVG' column).

sentences whose length ranges in an interval for which we have a reliable amount of instances (as introduced in Section 2.2).

Note that the different data altering strategies are conceived to represent increasingly challenging testbeds to assess the effectiveness of our probing tasks. The *Swapped* control datasets are the most challenging ones as the swapped feature values might be quite similar to the gold ones, thus possibly predicted with an high accuracy by the probing model. Such intuition is confirmed by the results of the 2-dimensional Principal Component Analysis (PCA) reported in Figure 1[3]. As we can see, all the data points representing the feature values contained in the *Swapped* datasets fully overlap with the gold ones, thus confirming their similarity. On the contrary, randomly generated values are progressively more distant being less plausible, even if the constraints of sentence length yield values that are closer to the gold ones.

3 Results

For both gold and control datasets, probing scores are computed as a Spearman correlation between the feature values predicted by the probing model and the values contained in each dataset. Such correlation values are computed by averaging the NLM's layer–wise scores as, for all datasets, we observed small differences between the scores obtained across the 12 layers. We experimentally verified that these differences were not significant by computing the slope of a linear regression line between BERT layers and the scores of the gold dataset, obtaining -0.0017 as mean value considering all features. Our intuition is that the small range of lengths of the sentences here considered may have yielded such insignificant variation across layers, which on the contrary Miaschi et al. (2021) showed to be significant on the whole set of IUDT sentences. Namely, being highly related to the length of the sentence, the feature values have little variations. However, a more in-depth investigation of the underlying reasons of this outcome is one of the future directions of this work.

Figure 2 shows the scores obtained on the gold and the 6 control datasets, both for the 7 macro-groups of linguistic features and on average (*AVG*). Additionally, in order to properly appreciate the differences between the results obtained on the *gold* and control datasets, in Figure 3 we report the error reduction rate for each control dataset computed as the difference between the scores obtained when predicting gold and altered features.

General Results. We can observe that on average the highest probing scores are obtained on the gold dataset and that, accordingly, there is a great difference (i.e. almost 1.0, see Figure 3) between

[3] PCA is a classical data analysis method that reduces the dimensionality of the data while retaining most of the variation in the data set by identifying *n* principal components, along which the variation of the data is maximal (Jolliffe and Cadima, 2016).

the accuracy of the probing model when predicting the authentic and altered feature values. This seems suggesting that the model is able to recognize that the feature values contained in the control datasets 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 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 scores obtained for the *Random* and *Swapped* datasets when we constrain the length of the sentences, we can observe that the accuracy in predicting the feature values contained in the *Swapped* datasets is sightly higher than in the *Random* ones (see 'AVG' column in Figure 2). This is in line with our starting hypothesis and shows that feature values artificially created simply by shuffling gold ones across sentences of the same lengths (or of the same range of lengths) are more similar to the gold values and thus are predicted with higher accuracy than randomly altered values. Nevertheless, their error rate, namely the difference from the accuracy of gold predictions, is still quite high, i.e. about 0.80 (see the 'AVG' column, Figure 3).

Linguistic Features Analysis. Also when we focus on the results obtained with respect to the 7 macro-groups of linguistic features, we can observe that the probing model is more accurate in the prediction of the gold values. Again, the scores on the control datasets are slightly higher when we constrain the values with respect to sentence length, since we narrow the range of possible values. In particular, we see that the feature values related to the sentence tree structure are those predicted most closely to the gold ones (see column 'TreeStructure', Figure 3). Note that these sentence properties are the most sensitive to the sentence length, that BERT encodes with a very high accuracy. This may suggest that in the resolution of these tasks the probing model is possibly relying on some regularities related to sentence length.

Dataset	Spearman correlation
Random	0.08
Random Bins	0.46 *
Random Lengths	0.33 *
Swapped	-0.15
Swapped Bins	0.05
Swapped Lengths	0.06

Table 2: Spearman correlations between the rankings of features obtained with the *Gold* dataset and the 6 control datasets. Statistically significant correlations are marked with * (p-value < 0.05).

Gold	Random Bins	Swapped Lengths
dep_dist_root	dep_dist_root	dep_dist_root
dep_dist_punct	avg_max_links_len	avg_max_links_len
upos_dist_PUNCT	max_links_len	max_links_len
xpos_dist_FS	xpos_dist_FB	avg_max_depth
upos_dist_ADP	avg_token_per_clause	verbal_head_per_sent
dep_dist_det	xpos_dist_FS	xpos_dist_FS
upos_dist_PROPN	n_prep_chains	avg_links_len
upos_dist_DET	avg_max_depth	subord_prop_dist
xpos_dist_RD	verbal_head_per_sent	avg_subord_chain_len
dep_dist_case	xpos_dist_RI	n_prep_chains
verbal_head_per_sent	dep_dist_cop	subord_post
xpos_dist_FF	xpos_dist_PC	subord_dist_1
xpos_dist_SP	dep_dist_conj	avg_prep_chain_len
xpos_dist_E	xpos_dist_B	obj_post
upos_dist_NOUN	xpos_dist_VA	avg_verb_edges

Table 3: 15 top-ranked *Gold* and control features (*Random Bins* and *Swapped Lengths*) predicted by BERT sentence-level representations.

Similar observations hold for the results achieved in the resolution of the probing tasks related to the use of subordination, which heavily depends on sentence length. Interestingly, we can note that the values of all the other groups of features contained in the control datasets are predicted by the probing model with a very low accuracy, possibly making the results not significant.

Features Correlations. Once we showed that the probing tasks accuracy is very different if the feature values are authentic or altered, in this section we compare the ranking of linguistic features ordered by decreasing prediction accuracy in the gold and control scenarios. As we can see in Table 2, which reports the Spearman correlations between the rankings, the *control rankings* are almost not related to the gold one and the existing correlations in most cases are not even statistically significant. The only exceptions are represented by the rankings of values that were randomly generated with sentence length constraints, which have a weak and moderate correlation. Note that however, as shown before, the probing scores are very low.

A more qualitative feature ranking analysis can be carried out by inspecting Table 3 where we report the first 15 top-ranked features predicted in the gold and in the two most highly correlated *Swapped* and *Random* datasets. As we can see, the *gold ranking* diverges from the rankings of the altered values with respect to the majority of top-ranked features. The most visible exception is represented by the distribution of syntactic root (*dep_dist_root*) that the probing model always predicts with the highest accuracy. The result is quite expected since this feature can be seen as a proxy of the length of the sentence, a linguistic property properly encoded by BERT. Similarly, other two features influenced by sentence length appear, as expected, on the top positions of all rankings, namely the distribution of the sentence boundary punctuation (*xpos_dist_FS*) and of verbal heads (*verbal_head_per_sent*).

4 Discussion and Conclusion

In this paper we described a methodology to test the effectiveness of a suite of probing tasks for evaluating the linguistic competence encoded by NLMs. To this aim, we analysed the performance of a probing model trained with BERT representations to predict the authentic and automatically altered values of a set of linguistic features derived from IUDT. We observed general higher performance in the prediction of authentic values, thus suggesting that the probing model relies on linguistic competences to predict linguistic properties. However, when we constrained automatically altered values with respect to sentence length, the model tends to learn surface patterns in the data.

As a general remark, it should be pointed out that our analyses dealt only with sentences showing a standard length (i.e., between 10 and 30 tokens per sentence). This choice, if on the one hand made our results more directly comparable across bins of sentences sharing the same length, on the other hand excluded from the analyses the shortest and the longest sentences of IUDT. Our future work will be devoted to replicate the probing task experiments described in this paper also on control datasets comprising sentences whose length is outside of the range considered here. To this aim, we performed preliminary analyses to test the scores of probing tasks on gold IUDT sentences that are less than 10-token and more than 30-token long. Interestingly, we noticed that the probing model is less accurate when predicting the linguistic features extracted from the group of IUDT short sentences. Specifically, the average Spearman correlation obtained on such group is 0.47, while probing scores on longer sentences (+30-token long) and on those used in our experiments achieved an average correlation of 0.56 and 0.66 respectively. Starting from this preliminary finding, a possible future investigation could focus on whether using longer or shorter sentences would also have an effect on the probing scores obtained with the control datasets.

In future work we also plan to investigate which features are more diagnostic of the linguistic competence encoded by a NLM and which ones, on the contrary, are more influenced by confounders, such as sentence length.

References

Yonatan Belinkov. 2021. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, pages 1–12, 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, May. 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, July. Association for Computational Linguistics.

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, November. 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, June. Association for Computational Linguistics.

Raffaele Guarasci, Stefano Silvestri, Giuseppe De Pietro, Hamido Fujita, and Massimo Esposito. 2021. Assessing BERT's ability to learn Italian syntax: a study on null-subject and agreement phenomena. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–15.

Rowan Hall Maudslay, Josef Valvoda, Tiago Pimentel, Adina Williams, and Ryan Cotterell. 2020. A tale of a probe and a parser. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7389–7395, Online, July. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word 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 4129–4138.

Ian T Jolliffe and Jorge Cadima. 2016. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202.

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, June. Association for Computational Linguistics.

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), December. International Committee on Computational Linguistics.

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

Joakim Nivre. 2015. Towards a universal grammar for natural language processing. In *Proceedings of The 16th Annual Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, pages 3–16. Springer.

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

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, April. Association for Computational Linguistics.

Jörg Tiedemann and Lars Nygaard. 2004. The OPUS corpus - parallel and free: `http://logos.uio.no/opus`. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May. European Language Resources Association (ELRA).

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

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, November. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Noëmi Aepli, Željko Agic, Lars Ahrenberg, et al. 2020. 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.*