Probing Linguistic Knowledge in Italian Neural Language Models across Language Varieties

Alessio Miaschi[*]
Università di Pisa;
ILC-CNR, Pisa - ItaliaNLP Lab

Gabriele Sarti[**]
CLCG, University of Groningen

Dominique Brunato[†]
ILC-CNR, Pisa - ItaliaNLP Lab

Felice Dell'Orletta[‡]
ILC-CNR, Pisa - ItaliaNLP Lab

Giulia Venturi[§]
ILC-CNR, Pisa - ItaliaNLP Lab

In this paper, we present an in-depth investigation of the linguistic knowledge encoded by the transformer models currently available for the Italian language. In particular, we investigate how the complexity of two different architectures of probing models affects the performance of the Transformers in encoding a wide spectrum of linguistic features. Moreover, we explore how this implicit knowledge varies according to different textual genres and language varieties.

1. Introduction and Motivation

In the last few years, the study of Neural Language Models (NLMs) and their representations has become a key research area in the Natural Language Processing (NLP) community. Several methods have been devised to obtain meaningful explanations regarding how these models are able to capture syntax- and semantic-sensitive phenomena (Belinkov and Glass 2019). Among them, the probing task approach has emerged as the most commonly adopted diagnostic strategy to estimate the mutual information shared by a neural network's parameters and some latent property that the model could have learned to encode in the training procedure. During probing experiments, a supervised model (*probe*) is trained to predict the latent information from the network's learned representations. If the probe does well, we may conclude that the network effectively encodes some knowledge related to the selected property. Formally speaking, let $f : x_i \rightarrow y_i$ be a neural network model mapping a corpus of input sentences $X = (x_1, \dots, x_n)$ to a set of target labels $Y = (y_1, \dots, y_n)$ for a learned downstream

[*] Department of Computer Science, Università di Pisa; Istituto di Linguistica Computazionale "Antonio Zampolli", CNR, Pisa - ItaliaNLP Lab. E-mail: ale.miaschi@gmail.com

[**] Center for Language and Cognition, University of Groningen. E-mail: g.sarti@rug.nl

[†] Istituto di Linguistica Computazionale "Antonio Zampolli", CNR, Pisa - ItaliaNLP Lab. E-mail: dominique.brunato@ilc.cnr.it

[‡] Istituto di Linguistica Computazionale "Antonio Zampolli", CNR, Pisa - ItaliaNLP Lab. E-mail: felice.dellorletta@ilc.cnr.it

[§] Istituto di Linguistica Computazionale "Antonio Zampolli", CNR, Pisa - ItaliaNLP Lab. E-mail: giulia.venturi@ilc.cnr.it

task. Assume that each sentence x_i is also labeled with some linguistic annotations z_i, reflecting the underlying properties we aim to detect. Let also $h_l(x_i)$ be the network's output at the l-th layer given the sentence x_i as input. To estimate the quality of representations h_l with respect to property z, a supervised model $g : h_l(x_i) \rightarrow z_i$ mapping representations to property values is trained. We take such model's performances as a proxy of $H(h_l(x), z)$. In information theoretic terms, the probe is trained to minimize entropy $H(z|h_l(x))$, and by doing that it maximizes mutual information between the two quantities.

(Alain and Bengio 2017) were among the first to use linear probing classifiers as tools to evaluate the presence of task-specific information inside neural networks' layers. The approach was later extended to the field of NLP by (Conneau et al. 2018) and (Zhang and Bowman 2018) *inter alia*, which evaluated the presence of semantic and syntactic information inside sentence embeddings generated by LSTM encoders (Hochreiter and Schmidhuber 1997) pretrained on different objectives using probing task suites.

Nowadays, several studies adopt the probing task approach to investigate the inner working of state-of-the-art Neural Language Models (NLMs). This approach demonstrated that NLMs representations encode linguistic knowledge in a hierarchical manner (Belinkov et al. 2017; Blevins, Levy, and Zettlemoyer 2018; Tenney et al. 2019), and can even support the extraction of dependency parse trees (Hewitt and Manning 2019). (Jawahar, Sagot, and Seddah 2019) investigated the representations learned by BERT (Devlin et al. 2019), one of the most prominent NLM, across its layers, showing that lower ones are usually better for capturing surface features, while embeddings from higher layers are better for syntactic and semantic properties. Using a suite of probing tasks, (Tenney, Das, and Pavlick 2019) deeply explore this behavior showing that the linguistic knowledge encoded by BERT through its 12/24 layers follows the traditional NLP pipeline.

While the vast majority of this research is focused on English contextual representations, relatively little work has been done to understand the inner working of non-English models. The study by (de Vries, van Cranenburgh, and Nissim 2020) represents an exception in this context: the authors applied the probing task approach to compare the linguistic competence encoded by a Dutch BERT-based model and multilingual BERT (mBERT), showing that earlier layers of mBERT are consistently more informative that earlier layers of the monolingual model. (Guarasci et al. 2021) applied instead the structural probe originally defined by (Hewitt and Manning 2019) on the representations of a pre-trained Italian BERT. Testing their approach on different subsets of the Italian Universal Dependency Treebank (IUDT), they showed on the one hand that the model is able to encode properties of syntax especially in its central-upper layers; on the other hand, that such embedded syntactic information can be used to successfully perform two specific syntactic tasks, i.e. prediction of Subject-Verb agreement and parsing of null-subject sentences. In (Guarasci et al. 2022), the authors exploited the same methodology to investigate the ability of multilingual BERT to transfer syntactic knowledge across the English, French and Italian languages.

Another less investigated issue in the previous studies has to do with the design of probing models themselves. Although many studies have focused on multiple transformer models and diagnostic tasks to probe their inner linguistic competence, few works tested different probing architectures and investigated more in-depth their actual effectiveness. Among this few works, (Hewitt and Liang 2019) were the first who observed 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. In addition, (Pimentel et al. 2020) showed that more complex probes, in contrast with simple linear models, could produce tighter estimates for the actual underlying information.

Starting from these premises, this paper introduces an approach to NLMs interpretation aimed at carrying out an in-depth investigation of the linguistic knowledge implicitly encoded by 6 Italian monolingual models and multilingual BERT. Besides the focus on Italian, which represents a scarcely considered language in the scenario of the NLM interpretation studies, a further novelty of our approach concerns the broad set of probing tasks we took into account, each corresponding to a specific property of sentence structure. In addition, the present study is one of the few that introduces a still rather under-investigated research issue, i.e. the comparative analysis of how and to which extent the different architectures on which the probing model rely on influence the probing accuracy. To address this point, for each Transformer, we perform the same suite of probing tasks using both a LinearSVR and a multilayer perceptron (MLP), and compare how each probing task's resolution is differently affected by the two architectures. Since all experiments were carried out on different sections of the Italian Universal Dependency Treebank (Zeman et al. 2019) considered as representative of different textual genres and language varieties, we are also able to investigate how linguistic knowledge of NLMs varies according to standard and less or non-standard varieties of the Italian language.

The present article is based on, and extends, the work reported in (Miaschi et al. 2020b).

Contributions. To the best of our knowledge, this is the first study aimed at comparing the linguistic knowledge encoded in the representations of multiple non-English pre-trained transformer models. In particular:

- we compare the probing performances of 7 pre-trained Italian NLMs spanning three models architectures over multiple linguistic features;
- we investigate how the complexity of the probing classifier impacts its ability to capture the information encoded in learned representations;
- we highlight how the implicit knowledge encoded by NLMs during the training process differs across textual genres and language varieties.

2. Approach

To inspect the inner knowledge of language encoded by the Italian Transformers, we relied on a suite of 82 probing tasks, each of which consists in predicting the value of a given feature modeling a specific linguistic property of the sentence. We tested two different probing architectures: a LinearSVR and a three-layer feedforward network with ReLU activations (Multi-layer perceptron, MLP). If the linear architecture is the most commonly used approach to infer information inside NLMs, the MLP was selected to investigate the presence of nonlinear relations in representations, which could hamper the probing performance of the LinearSVR probe. Regardless of the architecture, the two probing models take as input layer-wise sentence-level representations extracted from the Italian models. These representations are produced for each sentence of different sections of the Italian Universal Dependency Treebank (IUDT), version 2.5 (Zeman et al. 2019), and used to predict the actual value of each probing feature. Starting from the

Table 1
NLMs used in the experiments.

Name	Training data
BERT Architecture	
Multilingual-BERT	Wikipedia
BERT-base-italian	Wikipedia + OPUS (13GB)
AlBERTo	TWITA (191GB)
RoBERTa Architecture	
GilBERTo	OSCAR (71GB)
UmBERTo-Commoncrawl	OSCAR (69GB)
UmBERTo-Wikipedia	Wikipedia (7GB)
GPT-2 Architecture	
GePpeTto	Wikipedia + ItWAC (14GB)

results obtained we performed three complementary investigations. In the first one we compared the results obtained by the two probing architectures according to different groups of probing tasks (Section 3.1). Then, we move to compare the linguistic competence of the 7 Italian Transformers (Section 3.2). Finally, the impact of the considered linguistic varieties on the linguistic generalization abilities of the NLMs is discussed in Section 3.3.

2.1 Models and Data

We relied on 7 pre-trained Italian models based on three different Transformer architectures: BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019b) and GPT-2 (Radford et al. 2019). In particular, we investigated the linguistic competence of: three BERT-based models, i.e. Multilingual-BERT, BERT-base-italian[1] and AlBERTo (Polignano et al. 2019), trained respectively on Wikipedia (102 languages), Italian Wikipedia + texts from the OPUS corpus (Tiedemann and Nygaard 2004) and TWITA (Basile, Lai, and Sanguinetti 2018); three RoBERTa-based models, i.e. GilBERTo[2] and two versions of UmBERTo[3], trained respectively on OSCAR (Ortiz Suárez, Sagot, and Romary 2019) (GilBERTo and UmBERTo-Commoncrawl) and Italian Wikipedia; a GPT-2 based model, GePpeTto (De Mattei et al. 2020), trained on Italian Wikipedia + ItWAC (Baroni et al. 2009). Models statistics are reported in Table 1. Sentence level representations were computed performing a *Mean-pooling* operation over the word embeddings provided by the models across their layers.

NLM's linguistic competences are probed against 5 sections of the Italian Universal Dependency Treebank (IUDT) representative of different language varieties and textual genres, as shown in Table 2. The considered sections can be categorised in two main groups: a first one that includes sentences acquired from documents of diverse nature, ranging from Wikipedia pages, to newspaper articles, novels, speech transcriptions, etc., and a second group collecting examples of the social media language, in particular of Twitter. In the first group we included the Italian version of the multilingual Turin

[1] https://github.com/dbmdz/berts
[2] https://github.com/idb-ita/GilBERTo
[3] https://github.com/musixmatchresearch/umberto

Table 2
Sections of the Italian Universal Dependency Treebank (IUDT).

Short Name	Types of texts	# sent
ParTUT	Multi-genre	2,090
VIT	Multi-genre	10,087
ISDT	Multi-genre	14,167
ISDT_tanl	Newswire	4,043
ISDT_tut	Legal/Newswire/Wiki	3,802
ISDT_quest	Interrogative sentences	2,162
ISDT_2parole	Simplified Italian news	1,421
ISDT_europarl	EU Parliament debates	497
PoSTWITA	Tweets	6,713
TWITTIRÒ	Ironic Tweets	1,424
Total		35,481

University Parallel Treebank (ParTUT) (Sanguinetti and Bosco 2015), the Venice Italian Treebank (VIT) (Delmonte, Bristot, and Tonelli 2007) and Italian Stanford Dependency Treebank (ISDT) (Bosco, Montemagni, and Simi 2013), which we considered representative of the standard Italian language. The group of treebanks composed of PoSTWITA (Sanguinetti et al. 2018) and TWITTIRÒ (Cignarella, Bosco, and Rosso 2019) was originally built to enhance the performances of systems in processing social media texts, and in particular, for irony detection purposes. Being representative of a non-standard variety of the Italian language, for our specific scopes, they are intended to be a quite challenging testbed for probing the linguistic knowledge of NLMs also when they are trained on standard language variety.

Note that the linguistic abilities of the 7 NLMs were also tested against a number of sub-portions of the largest Italian UD treebank, i.e. ISDT. They have been chosen since they are representative of language sub-varieties possibly infrequently seen during the NLMs training phase. Accordingly, they can be conceived as a favorite point of view to investigate to which extent general-purpose NLMs are robust against less standard texts. For this purpose, in addition to sub-sections including newspapers (ISDT_tanl) and miscellaneous documents (ISDT_tut), we considered sub-portions including sentences in the interrogative form (ISDT_quest), newspaper articles specifically written to be linguistically simple (ISDT_2parole) and transcriptions of the European parlament oral debates (ISDT_europarl).

2.2 Probing features

The set of probing tasks consists in predicting the value of a specific linguistic feature automatically extracted from the manually revised annotation of each sentence of the IUDT datasets.

We relied on the set described in (Brunato et al. 2020) that includes about 130 features representative of the linguistic structure underlying a sentence and derived from raw, morpho-syntactic and syntactic levels of annotation. For the specific purpose of this study, we selected the 82 most frequent features in order to prevent data sparsity issues thus making our results reliable.

Table 3
Probing Features used in the experiments.

Linguistic Feature	**Label**
Raw Text Properties (*RawText*)	
Sentence Length	sent_length
Word Length	char_per_tok
Vocabulary Richness (*Vocabulary*)	
Type/Token Ratio for words and lemmas	ttr_form, ttr_lemma
Morphosyntactic information (*POS*)	
Distribution of UD and language–specific POS	upos_dist_*, xpos_dist_*
Lexical density	lexical_density
Inflectional morphology (*VerbInflection*)	
Inflectional morphology of lexical verbs and auxiliaries	verbs_*, aux_*
Verbal Predicate Structure (*VerbPredicate*)	
Distribution of verbal heads and verbal roots	verbal_head_dist, verbal_root_perc
Verb arity and distribution of verbs by arity	avg_verb_edges, verbal_arity_*
Global and Local Parsed Tree Structures (*TreeStructure*)	
Depth of the whole syntactic tree	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
Order of elements (*Order*)	
Relative order of subject and object	subj_pre, subj_post, obj_post
Syntactic Relations (*SyntacticDep*)	
Distribution of dependency relations	dep_dist_*
Use of Subordination (*Subord*)	
Distribution of subordinate clauses	subordinate_prop_dist
Average length of subordination chains and distribution by depth	avg_subord_chain_len, subordinate_dist_1
Relative order of subordinate clauses	subordinate_post

As shown in Table 3, the considered tasks are intended to probe whether the NLMs encode in their representations 9 main aspects of the structure of a sentence. They range from quite simple aspects related to the knowledge of raw text properties (i.e. sentence and word length), to the vocabulary richness (in terms of type/token ratio), to the distribution of UD and language-specific Parts-Of-Speech[4] and of inflectional properties specific in particular to verbal predicates (i.e. mood, tense, person). More challenging probing tasks concern the ability to encode complex aspects of sentence structure, including both global structure, such as the depth of the whole syntactic tree, and local features. We paid a specific attention to testing the models knowledge of the sub-trees of the nuclear elements of a sentence. In this respect, we included a group of features modelling the verbal predicate structure, e.g. in terms of number of dependents of verbal heads, and a group referring to the order of subjects and objects with respect to their verbal head. In line with the focus on specific sub-trees, we also considered a group

[4] For the list of UD Parts-Of-Speech refer to https://universaldependencies.org/u/pos/index.html, while for the language-specific one to http://www.italianlp.it/docs/ISST-TANL-POStagset.pdf

Table 4
Average R^2 scores for all the NLMs obtained with the LinearSVR and the MLP probing models. Baseline scores for a Linear SVR trained only on sentence length are also reported.

Groups	LinearSVR	MLP	Baseline
RawText	**0.84**	0.80	0.50
Vocabulary	**0.70**	0.34	0.19
POS	**0.69**	0.68	0.03
VerbInflection	0.50	**0.61**	0.03
VerbPredicate	0.32	**0.43**	0.08
TreeStructure	0.61	**0.64**	0.40
Order	0.46	**0.55**	0.06
SyntacticDep	0.65	**0.74**	0.04
Subord	0.49	**0.60**	0.16
AllFeatures	0.60	**0.64**	0.10

of features capturing the use of subordination in terms of distribution of subordinate clauses, of their internal structure and relative order with respect to the main clause.

We chose to rely on these linguistic characteristics 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, which guarantees the comparative encoding of language phenomena across different languages (Nivre 2015). In fact, they have been also used to profile the knowledge encoded in the language representations of a pretrained NLM, specifically the English BERT, and how it changes across layers (Miaschi et al. 2020a).

3. Experiments and Results

In this section we report the results of the three different investigations we carried out starting from the probing strategies devised.

3.1 Comparison of Probing Model Architectures

Our first analysis concerns the comparison of the two considered architectures for probing the linguistic knowledge encoded by the Italian Transformers. Since many of our probing features are strongly related to sentence length, we compared these results with the ones obtained by a baseline corresponding to a LinearSVR model trained using only sentence length as input feature. Table 4 reports average R^2 results[5] across all the layers of all the 7 NLMs obtained with the LinearSVR and the MLP probing architectures, along with baseline scores.

As a first remark, we notice that both probing architectures outperform the sentence length baseline. This suggests that all NLMs encode a spectrum of phenomena that,

[5] The Coefficient of determination (R^2) is a statistical measure of how close the data are to the fitted regression line and corresponds to the proportion of the variance in the dependent variable that is predictable from the independent variable(s).

although related to syntagmatic complexity, require a more sophisticated linguistic knowledge to be accurately predicted. However, if we compare the results achieved by the two architectures on all groups of linguistic phenomena (*AllFeatures*), we can see that MLP architecture achieves higher R^2 scores. This is specifically the case of the group of features which refer to characteristics of the verb inflectional morphology (*VerbInflection*) and structure (*VerbPredicate*) and the use of subordination (*Subord*), for which the differences between the two architectures is higher. On the contrary, the LinearSVR resulted to be more accurate to probe NLMs' competences of raw text properties, vocabulary richness and about the distribution of Parts-Of-Speech. Interestingly, the SVR architecture outperforms the MLP by more than .30 R^2 points when predicting features related to vocabulary richness (*Vocabulary*). The increase in performances observed for the MLP model on syntactic features can be motivated by the presence of nonlinearities in the probing model, which allow the model to capture non-linear relations between learned features. On the other hand, this increase in model capacity seems to hinder the performances of the probe on low level features (*RawText*, *Vocabulary*, *POS*) for which a simple linear combination can be sufficient. Despite this difference, a comparison of the rankings of linguistic phenomena ordered by decreasing scores for the two probing models shows that in both cases raw text properties and the distribution of morpho-syntactic categories (*POS*) appear in the first positions, while the order of subject and object (*Order*) and the structure of verbal predicates (*VerbPredicate*) are found in the lower part of the ranking. This observation suggests that the hierarchy of linguistic information captured by probing models is preserved, regardless of the architectural complexity of the probe. As a matter of fact, if we compute the Spearman correlation (ρ) between the average scores obtained for the 82 linguistic features with the LinearSVR and MLP we obtained a ρ of 0.71, thus indicating a strong correlations between the scores obtained with the two probing models.

In order to ensure that our probes are actually showing the linguistic generalization abilities of the NLMs rather than learning the linguistic tasks, we also tested the probing models using the *control task* approach devised in (Hewitt and Liang 2019). We produced a control version of the IUDT corpus by randomly shuffling the linguistic features assigned to each sentence and performed the same probing tasks with the two probing classifiers for all NLMs representations. The correlation and R^2 scores between regressors' predictions and shuffled scores were low (< 0.05) and comparable for both the SVR and the MLP. These results support the claim that NLMs representations encode information closely related to linguistic competence and that our probing models are not relying on spurious signals unrelated to our linguistic properties to solve the regression task.

3.2 Comparison of Italian Transformers

To investigate to which extent each transformer encodes the considered set of linguistic phenomena, we compared the performances achieved by the 7 NLMs, using the two probing architectures. Results are reported in Figure 1, where we can notice that the 7 Transformers achieve quite similar results when considering all features as a whole (*all*). Nevertheless, a more in depth analysis highlights a number of small differences. Namely, we can see that BERT-base-italian is the first and second best model for the MLP and SVR architecture respectively; while the least performing model is AlBERTo using MLP and, for the SVR probing architecture, UmBERTo model trained on the Italian Wikipedia.

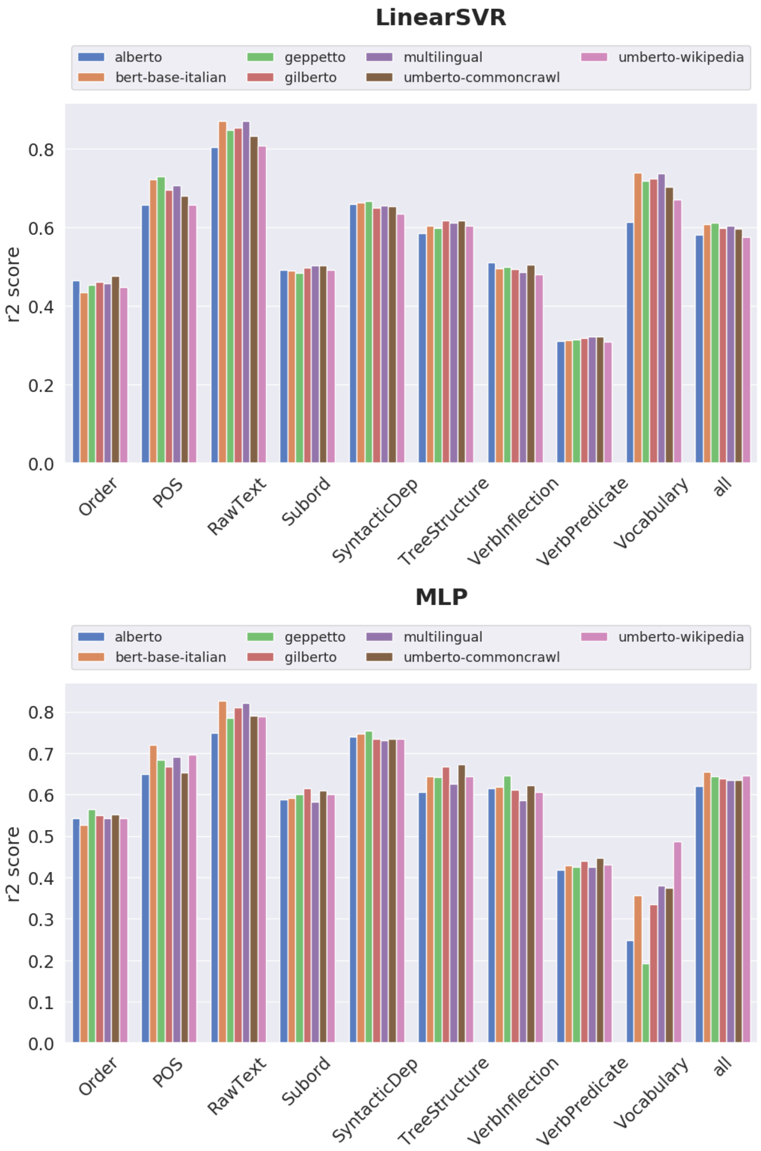

Figure 1
Layer-wise average R^2 scores obtained by each NLM with the two probing models.

However, this trend does not hold when we analyse the NLMs performances with respect to the encoding of the different groups of linguistic phenomena. For instance, we can notice that, for the two probing architectures, tree structure properties (*TreeStructure*) are predicted more accurately by RoBERTa-style models, i.e. by GilBERTo and UmBERTo-Commoncrawl, than by models based on BERT or GPT-2. Only for MLP, this can be similarly observed for the prediction of two other linguistic properties referring to sub-trees of the whole syntactic structure of a sentence. Namely, it can be seen that GilBERTo and UmBERTo-Commoncrawl are the two best models able to encode the use

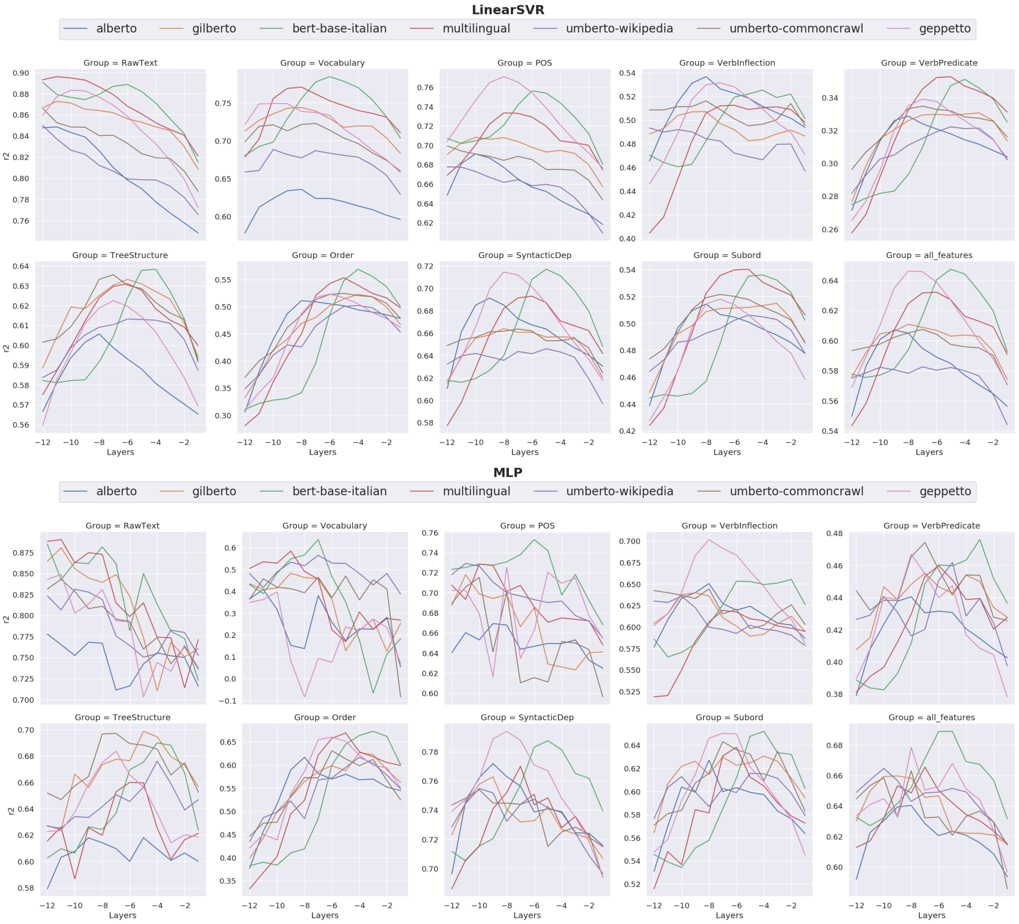

Figure 2
Average layerwise R^2 scores obtained with the LinearSVR (*top*) and the MLP (*bottom*) using the internal representations of the 7 NLMs.

of subordination (*Subord*) and the verb predicate structures (*VerbPredicate*). Further differences in terms of probing architectures can be inspected considering NLMs abilities to encode competencies related to vocabulary richness (*Vocabulary*): while UmBERTo-Wikipedia extensively outperforms all the other transformers using the MLP model, the best transformer is BERT-base-italian when these competences are probed with the LinearSVR model.

Additional observations can be made if we move to the analysis of how NLMs prediction abilities change and evolve across layers. As it can be seen in Figure 2, regardless of the architecture, for all transformers linguistic competences tend to decrease across the 12 layers. This is in line with previous findings (Liu et al. 2019a; Miaschi et al. 2020a) and it could be due to the fact that transformer layers trade off between task-oriented (e.g. Masked Language Modeling) information and general linguistic competence. Such decreasing trend can be specifically observed for example for the ability to predict raw text features, or the distribution of the UD morpho-syntactic categories (*POS*) and syntactic dependencies (*SyntacticDep*): they represent sentence properties mainly encoded in the first layers by all NLMs. On the contrary, we can observe that there is a number of more complex linguistic features whose knowledge increases consistently across layers, even if it decreases in the output layer. This is the case of features referring to structural sentence knowledge, such as the order of subject/object with respect to the verbal head (*Order*) and the use of subordination (*Subord*). In addition, contrarily to what was observed by (de Vries, van Cranenburgh, and Nissim 2020), Multilingual-BERT's linguistic knowledge is not encoded systematically earlier than in monolingual transformers.

This perspective of analysis also reveals other differences among the considered transformers which were unseen. By inspecting the trend of the R^2 scores across layers, we can for example see that even though GePpeTto has a lower average competence on verb inflection (see Figure 1), it achieves the highest scores in the middle layers. Or, even if we previously noted that RoBERTa-style transformers are more able to predict features related to the structure of a sentence (*TreeStructure*), the highest accuracy is achieved by a BERT-style model, i.e. BERT-base-italian, in the -4 layer. A similar observation also concerns the use of subordination and the verb predicate structure: the two groups of features are in general predicted more accurately by GilBERTo and UmBERTo-Commoncrawl but the highest R^2 scores are achieved by Mulilingual-BERT and BERT-base-italian in the -5 and -4 layers.

Focusing instead on differences between layerwise scores obtained by the two probing architectures, we can clearly notice that the encoding of linguistic knowledge shows a quite rough trend for what concerns the results obtained with the MLP. This is particularly the case of features belonging to the vocabulary, POS and tree structure groups.

If we deepen our investigation and we focus on the linguistic generalization ability of the NLMs with respect to each individual feature (see Figure 3), we can clearly observe that the rankings according to R^2 scores are quite similar regardless the probing architecture and the transformer model. It is also interesting to note that, despite some deviations, the distinction into macro-groups of linguistic phenomena seems to be mostly preserved across the rankings. In fact, raw-text features, as well as the distributions of POS-tags (*upos_dist_**, *xpos_dist_**) and dependency relations (*dep_dist_**), are those that were better predicted by the two probing models, while features more related to the structural information of a sentence, such as the order of elements (e.g. *subj_pre*, *subj_post* and *obj_post*) or the structure of parsed tree (e.g. *avg_token_per_clause*, *avg_prep_chain_len*) achieved lower probing scores. Lower results also concern the prediction of the morphological features of lexical and auxiliary verbs, namely for example their mood (*verb_mood_**) or tense (*verb_tense_**).

In line with what observed in Figure 1, we can see that in few cases the linguistic competence of the AlBERTo model is significantly different (lower) from that of the other models. The most remarkable case concerns the distribution of punctuation marks in general, both at the level of morpho-syntactic category (*upos_dist_PUNCT*), dependency relation (*dep_dist_punct*), and more specifically considering the distribution of commas (*xpos_dist_FF*) and balanced punctuation (*xpos_dist_FB*). This appears particularly evident using MLP as probing architecture and it is possibly related to the typology of texts the AlBERTo model was trained on, i.e. Twitter. It is well known that social media represents a non standard language variety, characterised by specific linguistic properties mostly different from ordinary language (Farzindar and Inkpen 2015), such as short sentences where punctuation marks, especially weak ones, are rarely used. Accordingly, the low frequency of punctuation in the training corpus possibly yields AlBERTo's reduced generalization abilities with respect to this specific set of features.

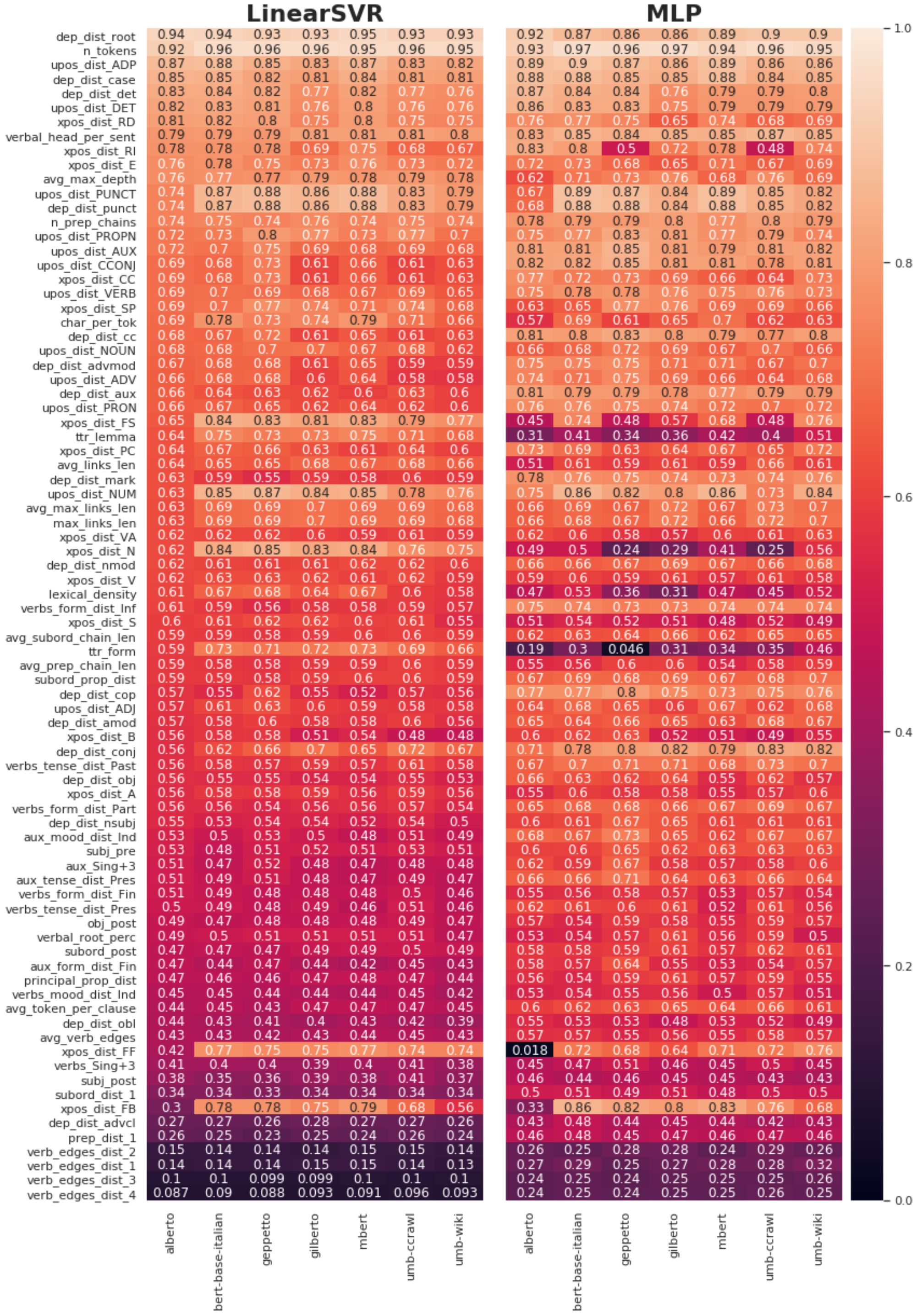

Figure 3
Average R^2 scores obtained for each probing features using the two probing architectures tested with the internal representations of the 7 NLMs. Both heatmaps are ordered on the basis of the feature ranking as predicted by the AlBERTo model using the LinearSVR architecture.

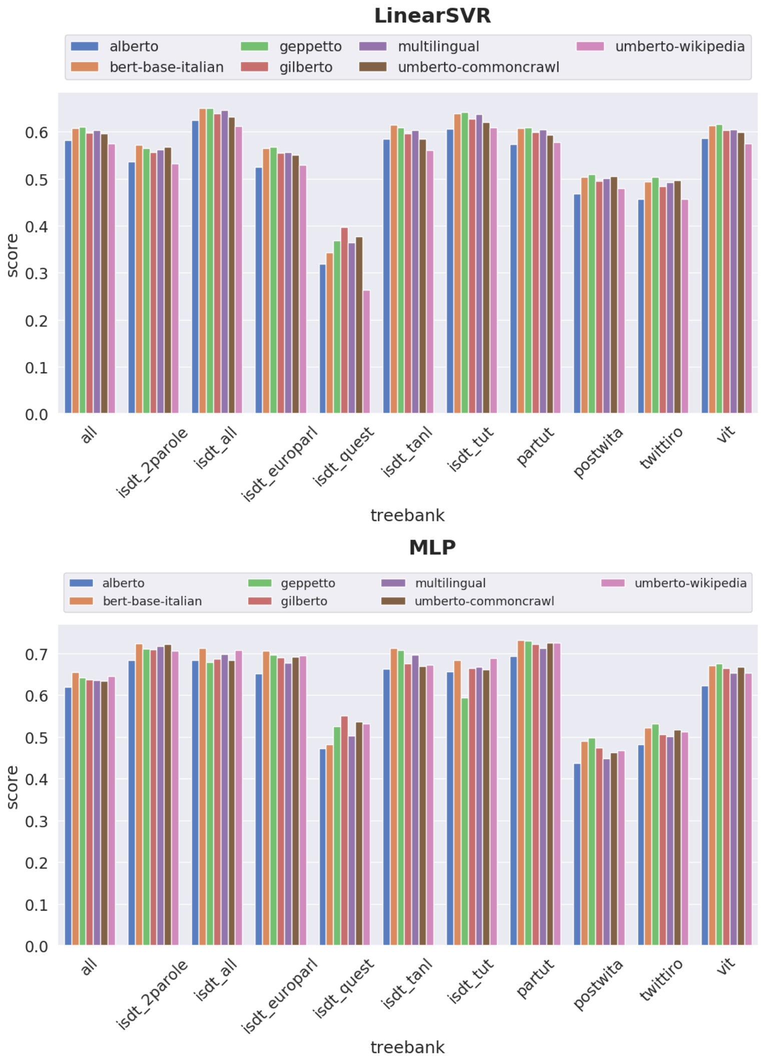

Figure 4
Average LinearSVM R^2 score considering all the UD Italian sentences (*all*) and according to the 10 treebanks previously described.

3.3 Comparison of Italian Language Varieties

Our last analysis concerns the impact of the considered Italian language varieties on NLMs linguistic abilities. For this purpose, we inspected whether the overall linguistic competence encoded in the contextual representations of each model changes according to the different IUDT sections. The results reported in Figure 4 show that all transformers, regardless of the probing architecture, achieve lower performance when they have to predict the value of features extracted from treebanks representative of social media language (PoSTWITA and TWITTIRÒ) and from the sub-set of ISDT sentences in the interrogative form (ISDT_quest). In both cases, this seems supporting our starting intuition that NLMs trained on standard language varieties, represented for example by Wikipedia pages, websites or web-crawled documents, may be less robust to non-standard varieties that were possibly unseen, or rarely seen, during the pre-training process. Quite surprisingly, even if AlBERTo has been trained on Twitter data, it obtains the lowest R^2 scores also when its internal representations are used to predict the feature values of the two social media Italian treebanks. A possible explanation is that, although PoSTWITA and TWITTIRÒ contain sentences representative of Twitter language, these sentences are still quite close to the Italian standard language, in order to be compliant with the UD morpho-syntactic and syntactic annotation schema. On the contrary, AlBERTo's training set is derived from Twitter's official streaming API that included all possible typologies of sentences.

It also worth noting that BERT-base-italian and GePpeTto are the two models slightly less affected by the non-standard linguistic peculiarities of the social media variety. As noted in Section 3.2, they represent the two best performing models in terms of overall linguistic competence. This may explain why they are more robust in the accurate prediction of the features values of all the considered IUDT sections. This holds both with the LinearSVR and MLP probing architecture, even if in the latter case the two versions of UmBERTo achieve comparable or slightly better scores. A main exception is represented by the ISDT sub-section including sentences in the interrogative form (ISDT_quest), which, as we noted above, are hardly mastered by all models. This is possible due to the fact that interrogative sentences are more likely to display a less canonical distribution of morpho-syntactic and syntactic phenomena, hence being more difficult to encode effectively. In this case, the transformer based on GPT-2, i.e. GePpeTto, results to be the NLM with the highest linguistic knowledge of this type of sentences.

A further analysis of the impact of language varieties on the ability of NLMs to encode the considered group of linguistic phenomena can be appreciated in Table 5. It shows, for each probing architecture, the Spearman correlations between the rankings of features predicted by all NLMs considering three ISDT sub-sections, i.e. ISDT_tanl, ISDT_2parole and ISDT_quest, and PoSTWITA, and ordered by decreasing R^2 scores. For each NLM, higher correlations correspond to similar linguistic generalization abilities across the paired treebanks, while lower correlations suggest that the inner representations of the NLM allow predicting effectively diverse linguistic features. As we can see, regardless of the probing architecture, for all NLMs, the highest correlated rankings are those obtained comparing ISDT_tanl (*tanl*) and PoSTWITA (*ptw*) predicted features. Even if it is quite surprising, this result can be explained assuming that the morpho-syntactic and syntactic features of the Twitter sentences contained in PoSTWITA are not so dramatically different from those characterising ISDT_tanl newspaper articles. In fact, among all the IUDT sections considered here they resulted to be the two most similar treebanks with respect to the distribution of the set of linguistic features reported in Table 3. In particular, the main differences concern the distribution of some morpho-syntactic categories (e.g. punctuation, nouns) and main features related to the inflectional morphology of verbs, e.g. the distribution of present tenses, higher in PoSTWITA (51.11% out of the total verb tenses) than in ISDT_tanl (34.95%), or of the past tenses that in the Twitter sentences are less than half than in the newspaper ones. Interestingly, these characteristics belong to the group of features that the NLMs are able to master quite accurately, regardless of the language variety. Even if these differences had a negative impact on the overall probing abilities of the PoSTWITA sentence characteristics, as shown in Figure 4, the higher knowledge of these specific features did not possibly

Table 5
Spearman correlations between rankings of features as predicted by the 7 NLMs on four sections of the IUDT treebank: IUDT_2parole (*2par*), IUDT_tanl (*tanl*), IUDT_quest (*quest*) and IUDT_postwita (*ptw*). Highest correlations are bolded, while lowest ones are marked in italics.

Model	**Section**	**LinearSVR**				**MLP**			
		2par	tanl	quest	ptw	2parole	tanl	quest	ptw
alberto	2par	1				1			
	tanl	.72	1			**.85**	1		
	quest	*.38*	*.38*	1		.62	*.56*	1	
	ptw	.76	**.82**	.45	1	.75	.80	.58	1
bert-base-italian	2par	1				1			
	tanl	.68	1			.82	1		
	quest	*.34*	.41	1		.62	*.47*	1	
	ptw	.72	**.91**	.47	1	.75	**.88**	*.47*	1
geppetto	2par	1				1			
	tanl	.65	1			.80	1		
	quest	*.30*	.38	1		.64	50	1	
	ptw	.70	**.92**	.48	1	.72	**.88**	*.47*	1
gilberto	2par	1				1			
	tanl	.61	1			.77	1		
	quest	*.30*	.40	1		.58	54	1	
	ptw	.66	**.88**	.46	1	.69	**.82**	*.49*	1
mbert	2par	1				1			
	tanl	.65	1			.76	1		
	quest	*.30*	.37	1		.55	.47	1	
	ptw	.71	**.90**	.45	1	.71	**.83**	*.46*	1
umberto-commoncr.	2par	1				1			
	tanl	.58	1			.71	1		
	quest	*.28*	.33	1		.55	.47	1	
	ptw	.69	**.8**	.39	1	.65	**.75**	*.35*	1
umberto-wikipedia	2par	1				1			
	tanl	.57	1			.70	1		
	quest	-	-	1		.50	.44	1	
	ptw	.66	**.72**	*.36*	1	.69	**.72**	*.36*	1

have a great consequence on the ranking of the predicted features, thus yielding quite high correlations.

On the contrary, the lowest correlations can be observed when we compare the rankings obtained for the pairs of treebanks containing the set of sentences in the interrogative form, i.e. ISDT_quest (*quest*). Even if the correlation values are slightly higher using MLP, this trend holds for the two probing architectures and for all NLMs. Note that the correlations between the ranking obtained with UmBERTo-Wikipedia for the pairs ISDT_quest/ISDT_2parole and ISDT_quest/ISDT_tanl are even not statistically significant. Let us remind that this is the NLM that achieved the lowest prediction accuracy using the LinearSVR probing architecture (see Figure 1). Our intuition is that this may have made it less robust in the prediction of non-standard linguistic forms, such as interrogative sentences. Similarly to what aforementioned, these results can be explained if we analyse the feature values in the considered treebanks. ISDT_quest resulted to be quite different from all the other treebanks particularly with respect to complex aspects of sentence structure. For example, the canonical order of the nuclear elements of a sentence (i.e. subject and object) is largely subverted in sentences in the interrogative form. Thus, they contain a very high percentage of post-verbal explicit subjects (68.69% of the total), half an order of magnitude higher than ISDT_tanl (15.21%) and PoSTWITA (12.63%) and an order of magnitude higher than ISDT_2parole (7.55%). Sentences in the interrogative form also have a lower percentage of post-verbal objects (17.31%), which instead represent the majority of cases in other treebanks, and they are characterised by a very low distribution of subordinate clauses in general and in particular of subordinates following the principal clause, i.e. 4% vs. 43% in ISDT_tanl, 35.78% ISDT_2parole and 44.36%. These and other similar features all concern structural aspects of a sentence that may have undermined the overall NLM linguistic competence thus yielding not only lower probing scores on ISDT_quest but also different feature rankings with respect to the other treebanks.

4. Conclusion

In this paper we presented an in-depth comparative investigation of the linguistic knowledge encoded in the Italian transformer models. Relying on a suite of 82 probing features and on two different probing architectures, we performed a number of complementary investigations all tested on different sections of the Italian Universal Dependency Treebank (IUDT), representative of diverse textual genres and language varieties.

Firstly, we showed experimentally how non-linear architectures such as the multilayer perceptron (MLP) capture a broader range of information encoded in learned representations with respect to their linear counterparts, and as such they can be considered more suitable for studying highly nonlinear models such as NLM. In this sense, our results support the information-theoretic operationalization of probing proposed by (Pimentel et al. 2020). However, the rankings of this and of the LinearSVR model in terms of their probing ability are quite similar. Namely, both are particularly able to probe raw text properties, as well as the distribution of Parts-Of-Speech and dependency relations; while they obtained lower scores for features referring to the order of subject and object with respect their verbal head and to the verbal predicate structure.

The following comparison of the linguistic generalization abilities of the 7 Transformers showed that if we analyse the results considering all the probing features as a whole few differences can be observed. Similarly to what observed for English (Liu et al. 2019a) and Dutch (de Vries, van Cranenburgh, and Nissim 2020), we showed that regardless of the probing architecture, for all transformers the internal layers (i.e. -6/-4) are the most informative ones and the linguistic competences tend to decrease across the 12 layers. However, contrary to (de Vries, van Cranenburgh, and Nissim 2020) our findings reveal that Multilingual-BERT's linguistic knowledge is not encoded systematically earlier than in monolingual transformers. More interesting outcomes result when we focus on the embedded knowledge of each group of linguistic characteristics. We noticed for example that global and local tree structure properties are predicted more accurately by RoBERTa-style models, i.e. by GilBERTo and UmBERTo-Commoncrawl, than by models based on BERT or GPT-2. We obtained additional information when we narrowed our analysis on how NLMs prediction abilities evolve across models' layers, showing for example that the highest competence about the tree structure is achieved by a BERT-style model, i.e. BERT-base-italian, in the -4 layer. A more in-depth comparison with respect to the ranking of each individual feature by R^2 scores also revealed that, even if the 7 Transfomers are quite similar, a main exception is represented by the AlBERTo model. In particular, it showed to have reduced generalization abilities concerning the use of punctuation. Our intuition is that it is possibly related to the typology of texts the AlBERTo model was trained on, i.e. Twitter, where punctuation marks are rarely used.

Finally, we showed that the level of NLMs linguistic competence changes according to the diverse linguistic varieties of IUDT. All Transformers resulted to be less robust in the prediction of the linguistic properties characterising sentences representative of social media language and of sentences in the interrogative form. This is possible due to the fact that the two types of sentences are characterised by non-canonical distribution of morpho-syntactic and syntactic phenomena, possibly rarely or never seen during the training phase. Surprisingly, also the AlBERTo model, even if it was trained on Twitter data, achieved very low performances, while on the contrary, BERT-base-italian and GePpeTto are the two models slightly less affected by the non-standard linguistic varieties. Despite both social media and questions seem representing two quite challenging testbeds, our in-depth investigation of how each probing feature is ranked by the NLMs allowed highlighting noteworthy differences. We observed that the most diverse rankings concern the test on the sentences in the interrogative form, which result to be characterised by distributions of structural aspects very different from other IUDT sections.

In terms of present and future research directions, we are currently investigating how the relation between the linguistic knowledge encoded by a NLM positively affects the resolution of downstream tasks, following recent works highlighting the tendency of pretrained NLMs to lose general linguistic information during the fine-tuning process and the connection between encoded linguistic information and models' downstream performances for the English language (Miaschi et al. 2020a; Sarti, Brunato, and Dell'Orletta 2021). These connections, which are still sporadically investigated at the moment, can cast a light on the decision process inside NLMs, and ultimately lead to an improved understanding and utilization of these systems for real-world usage.

References

Alain, Guillaume and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. In *Workshop Track of the Fifth International Conference on Learning Representations (ICLR 2017)*, Toulon, France, April.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Basile, Valerio, Mirko Lai, and Manuela Sanguinetti. 2018. Long-term social media data collection at the university of Turin. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, volume 2263, pages 1–6, Turin, Italy, December. CEUR Workshop Proceedings.

Belinkov, Yonatan and James Glass. 2019. Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, 7:49–72, 04.

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

Blevins, Terra, Omer Levy, and Luke Zettlemoyer. 2018. Deep RNNs encode soft hierarchical syntax. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–19, Melbourne, Australia, July. Association for Computational Linguistics.

Bosco, Cristina, Simonetta Montemagni, and Maria Simi. 2013. Converting Italian treebanks: Towards an Italian Stanford dependency treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 61–69, Sofia, Bulgaria, August. Association for Computational Linguistics.

Brunato, Dominique, 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.

Cignarella, Alessandra Teresa, Cristina Bosco, and Paolo Rosso. 2019. Presenting TWITTIRÒ-UD: An Italian Twitter treebank in Universal Dependencies. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 190–197, Paris, France, August. Association for Computational Linguistics.

Conneau, Alexis, 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.

De Mattei, Lorenzo, Michele Cafagna, Felice Dell'Orletta, Malvina Nissim, and Marco Guerini. 2020. GePpeTto carves italian into a language model. In Felice Dell'Orletta, Johanna Monti, and Fabio Tamburini, editors, *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, pages 136–143, Bologna, Italy (Online), March.

de Vries, Wietse, 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.

Delmonte, Rodolfo, Antonella Bristot, and Sara Tonelli. 2007. VIT - Venice Italian Treebank: Syntactic and quantitative features. In *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories*, Bergen, Norway, August.

Devlin, Jacob, 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.

Farzindar, Atefeh and Diana Inkpen. 2015. *Natural Language Processing for Social Media*. Synthesis Lectures on Human Language Technologies, Morgan & Claypool.

Guarasci, Raffaele, 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.

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

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

Hewitt, John 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, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Hochreiter, Sepp and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jawahar, Ganesh, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy, July. Association for Computational Linguistics.

Liu, Nelson F., Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. 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.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Miaschi, Alessio, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2020a. 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.

Miaschi, Alessio, Gabriele Sarti, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2020b. 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 2021. CEUR.org.

Nivre, Joakim. 2015. Towards a universal grammar for natural language processing. In Alexander Gelbukh, editor, *Computational linguistics and intelligent text processing*, pages 3–16, New York. Springer.

Ortiz Suárez, Pedro Javier, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In Piotr Bański, Adrien Barbaresi, Hanno Biber, Evelyn Breiteneder, Simon Clematide, Marc Kupietz, Harald Lüngen, and Caroline Iliadi, editors, *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019*, pages 9 – 16, Cardiff, 22nd July. Leibniz-Institut für Deutsche Sprache.

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

Polignano, Marco, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. 2019. Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets. In Raffaella Bernardi, Roberto Navigli, and Giovanni Semeraro, editors, *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, Bari, Italy, November.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *Technical report*.

Sanguinetti, Manuela and Cristina Bosco. 2015. PartTUT: The turin university parallel treebank. In Roberto Basili et al., editor, *Harmonization and Development of Resources and Tools for Italian Natural Language Processing within the PARLI Project*. Springer, pages 51–69.

Sanguinetti, Manuela, Cristina Bosco, Alberto Lavelli, Alessandro Mazzei, Oronzo Antonelli, and Fabio Tamburini. 2018. PoSTWITA-UD: an Italian Twitter treebank in Universal Dependencies. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

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

Tenney, Ian, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy, July. Association for Computational Linguistics.

Tenney, Ian, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, et al. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *Proceedings of the Seventh International Conference on Learning Representations (ICLR 2019)*, New Orleans, Louisiana, USA, May.

Tiedemann, Jörg 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).

Zeman, Daniel, Joakim Nivre, Mitchell Abrams, and al. 2019. Universal dependencies 2.5. In *LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL)*.

Zhang, Kelly and Samuel Bowman. 2018. Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. In *Proceedings of the 2018*

EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 359–361, Brussels, Belgium, November. Association for Computational Linguistics.