Italian Transformers Under the Linguistic Lens

Alessio Miaschi[♣,★], **Gabriele Sarti**[♠,♦,★], **Dominique Brunato**[★], **Felice Dell'Orletta**[★], **Giulia Venturi**[★]

[♣]Department of Computer Science, University of Pisa

[♠]Department of Mathematics and Geosciences, University of Trieste

[♦]International School for Advanced Studies (SISSA), Trieste

[★]Istituto di Linguistica Computazionale "Antonio Zampolli", ItaliaNLP Lab, Pisa

alessio.miaschi@phd.unipi.it, gsarti@sissa.it,

{name.surname}@ilc.cnr.it

Abstract

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 whether and how using different architectures of probing models affects the performance of Italian transformers in encoding a wide spectrum of linguistic features. Moreover, we explore how this implicit knowledge varies according to different textual genres.

1 Introduction and Background

In the last few years, the study of Neural Language Models (NLMs) and their representations has become a key research area in the NLP community. Several methods have been devised to obtain meaningful explanations regarding the linguistic information encoded in NLMs (Belinkov and Glass, 2019). The most common approach is based on the development of *probes*, i.e. supervised models trained to predict a variety of language properties using the contextual word/sentence embeddings of a pre-trained model (Conneau et al., 2018; Zhang and Bowman, 2018; Miaschi and Dell'Orletta, 2020). This approach demonstrated that NLMs representations encode linguistic knowledge in a hierarchical manner (Belinkov et al., 2017; Blevins et al., 2018; Tenney et al., 2019b), and can even support the extraction of dependency parse trees (Hewitt and Manning, 2019). Jawahar et al. (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 et al. (2019a) 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 focused on English contextual representations, relatively little work has been done to understand the inner workings of non-English models. The study by de Vries et al. (2020) represents an exception in this context: authors apply 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. The survey by Nozza et al. (2020) also provides a comparative study of mBERT and language-specific BERT models but focused on the performance that each model obtains after training on several specific downstream tasks.

In this paper, we adopt a task-agnostic perspective to carry out an in-depth investigation of the linguistic knowledge implicitly encoded by 6 Italian monolingual models and multilingual BERT. We define a broad set of probing tasks, each corresponding to a specific property of sentence structure. We then compare the average performance reached by each model in predicting the feature value, evaluating the results obtained by models using their layer-wise sentence-level representations. A further comparative perspective, which to our knowledge is still rather under-investigated, concerns the study of how the architecture of the probing model itself influences probing scores. To address this point, for each model, we perform the same suite of probing tasks using both a linear SVR and a multilayer perceptron (MLP), and compare whether and how each probing task's resolution is affected by the two architectures.

Since all experiments were carried out on different sections of Italian Universal Dependency Treebank (Nivre et al., 2016), we were also able to investigate how linguistic knowledge of NLMs varies according to different textual genres.

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: (i) we compare the probing performances of 6 Italian NLMs spanning three models over multiple linguistic feature categories; (ii) we investigate whether and how using different architectures of probing models affects the performance of transformers in encoding specific features; and (iii) we show how the implicit knowledge learned by these models differs across textual genres.

2 Approach

To inspect the inner knowledge of language encoded by Italian Transformers, we relied on a suite of 82 probing tasks, each of which corresponds to predicting the value of a corresponding feature modeling a specific property of the sentence. We designed two sets of experiments. The first one consists in comparing the linguistic knowledge encoded by the Italian Transformers and evaluating the best probing model for inferring such knowledge from the NLMs. We compared the results obtained with two simple probing models, a linear SVR and a multilayer perceptron (MLP), which 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. In the second set of experiments, we evaluated how the Italian models' linguistic knowledge differs across textual genres and varieties, considering different IUDT sections.

2.1 Models and Data

We relied on 7 pre-trained Italian Transformers models. Models statistics are reported in Table 1.

[1]https://github.com/dbmdz/berts
[2]Polignano et al. (2019)
[3]https://github.com/idb-ita/GilBERTo
[4]https://github.com/musixmatchresearch/umberto
[5]De Mattei et al. (2020)

Name	Training data
BERT Architecture	
Multilingual-BERT	Wikipedia
BERT-base-italian[1]	Wikipedia + OPUS (13GB) (Tiedemann and Nygaard, 2004)
AlBERTo[2]	TWITA (191GB) (Basile et al., 2018)
RoBERTa Architecture	
GilBERTo[3]	OSCAR (71GB) (Suárez et al., 2019)
UmBERTo-Commoncrawl	OSCAR (69GB)
UmBERTo-Wikipedia[4]	Wikipedia (7GB)
GPT-2 Architecture	
GePpeTto[5]	Wikipedia + ItWAC (14GB) (Baroni et al., 2009)

Table 1: NLMs used in the experiments.

Short Name	Types of texts	# sent
ParTUT (Sanguinetti and Bosco, 2015)	Multi-genre	2,090
VIT (Delmonte et al., 2007)	Multi-genre	10,087
ISDT (Bosco et al., 2013)	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 acts	497
PoSTWITA (Sanguinetti et al., 2018)	Tweets	6,713
TWITTIRÒ (Cignarella et al., 2019)	Ironic Tweets	1,424
Total		35,481

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

Sentence level representations were computed performing a *Mean-pooling* operation over the word embeddings provided by the models.

NLM's linguistic competences are probed against five IUDT sections including texts representative of different textual varieties and genres. As shown in the overview in Table 2, we also distinguish the whole ISDT into different sub–corpora according to the specific language variety they represent, e.g. transcription of spontaneous speech (*ISDT_europarl*), questions (*ISDT_quest*) or simplified language (*ISDT_2parole*).

2.2 Probing features

The set of probing tasks consists of predicting the value of a specific linguistic feature automatically extracted from each POS tagged and dependency parsed sentence of the IUTD datasets.

The set of features is based on the ones described in Brunato et al. (2020) and are acquired from raw, morpho-syntactic and syntactic levels of annotation and can be categorised in 9 groups corresponding to different linguistic phenomena. As shown in Table 3, these features model linguistic phenomena ranging from raw text one, to morpho–syntactic information and inflectional properties of verbs, to more complex aspects of sentence struc-

Linguistic Feature
Raw Text Properties
Sentence Length
Word Length
Vocabulary Richness
Type/Token Ratio for words and lemmas
Morphosyntactic information
Distibution of UD and language–specific POS
Lexical density
Inflectional morphology
Inflectional morphology of lexical verbs and auxiliaries
Verbal Predicate Structure
Distribution of verbal heads and verbal roots
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
Clause length
Relative order of elements
Order of subject and object
Syntactic Relations
Distribution of dependency relations
Use of Subordination
Distribution of subordinate and principal clauses
Average length of subordination chains and distribution by depth
Relative order of subordinate clauses

Table 3: Probing Features used in the experiments.

ture capturing global and local properties of the whole parsed tree and of specific subtrees, such as the order of subjects and objects with respect to the verb, the distribution of UD syntactic relations, also including features referring to the use of subordination and to the structure of verbal predicates.

All these features have been shown to play a highly predictive role when leveraged by traditional learning models on a variety of classification problems, covering different aspects of stylometric and complexity analysis. In addition, in their recent work, Miaschi et al. (2020) showed that these features can be effectively used to profile the knowledge encoded in the language representations of a pretrained NLM, specifically the English Bert, and how it changes across layers. Since these features are based on the UD formalism, which guarantees the comparative encoding of language phenomena between the two languages (Nivre, 2015), we focused on the same set to investigate the linguistic knowledge of Italian transformers.

3 Results

We first investigate which is the best architecture for probing the linguistic knowledge encoded by the Italian Transformers. Since many of our probing features are strongly related to sentence length, we compared the two probing models' 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[6] for all the 7 NLMs obtained with the LinearSVR and the MLP probing models, along with baseline scores. The MLP probe is a three-layer feedforward network with ReLU activations and was selected to investigate the presence of nonlinear relations in representations, which could hamper the probing performance of the LinearSVM probe, but would be highlighted by a sharp difference between MLP and LinearSVM performances. As a first remark, we notice that both probing models outperform the baseline. This proves that all NLMs encode a spectrum of phenomena that, although related to sentence length, require a more sophisticated linguistic knowledge to be accurately predicted. Best scores are obtained with the MLP model, which achieved higher R^2 scores especially for features grouping more complex syntactic phenomena (e.g. *TreeStructure*, *SyntacticDep*). Interestingly enough, the LinearSVR model outperforms the MLP by more than .30 R^2 points when predicting features related to vocabulary richness (*Vocabulary*).

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

Table 4: Average R^2 scores for all the NLMs obtained with the LinearSVR and the MLP probing models. Baseline scores are also reported.

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 correla-

[6]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).

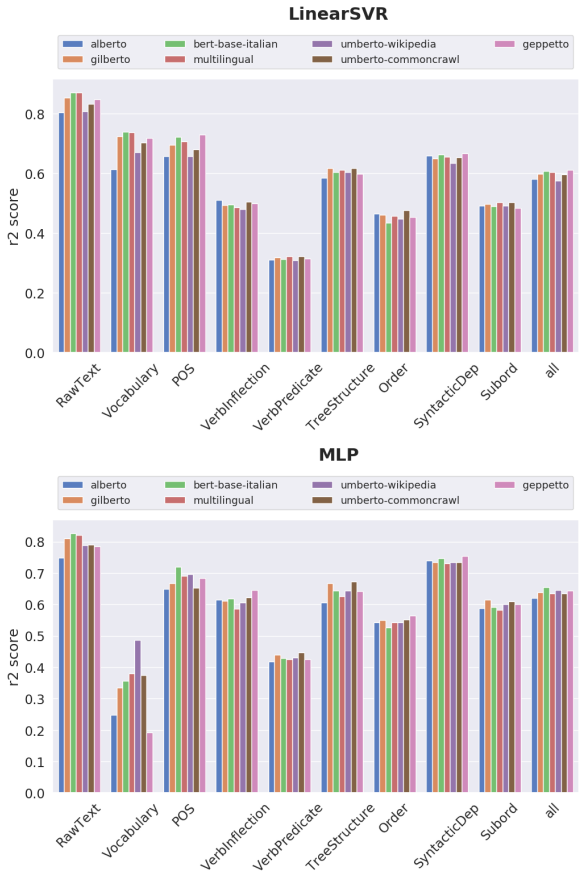

Figure 1: Average R^2 scores obtained by each NLM with the two probing models.

tion 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.

To investigate how each transformer encodes the linguistic knowledge, we report in Figure 1 average R^2 scores obtained with the two probing models for all the 7 NLMs. As we can notice, the seven transformers achieve quite similar results when considering all features as a whole, although BERT-base-italian has the best overall performance (0.65 for *all* features). The same did not hold when we analyzed their performances in terms of R^2 scores for the different previously described groups of features. For instance, we can notice that, for both the probing models, features related to the distribution of syntactic relations (*SyntacticDep*) are better predicted by GePpeTto, while GilBERTo and UmBERTo-Commoncrawl are the best ones in the prediction of tree structure properites. Differences hold for what regards 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.

Similar trends can be observed in Figure 2, where we report how the linguistic knowledge encoded by the 7 NLMs evolves across layers according to the two probing models. Regardless of the architectures, for all transformers, raw text features (*RawText*) are mainly encoded in the first layers, while the knowledge about the order of subject/object (*Order*) and the use of subordination (*Subord*) increases consistently across layers and specifically in the first ones. Contrarily to what was observed by de Vries et al. (2020), mBERT's linguistic knowledge is not encoded systematically earlier than in monolingual transformers. This perspective of analysis also reveals other differences among the considered transformers: e.g. even though GePpeTto has a lower average competence on verb inflection (see Figure 1), it achieves the highest scores in the middle layers. Focusing instead on differences between layerwise scores obtained by the two probing models, 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.

Finally, we inspected whether the overall linguistic competence encoded in the contextual representations of each model changes according to the type of texts in the different IUDT sections we considered. As we could expect, the results reported in Figure 3 show that all transformers achieve lower performance when they have to predict the value of features extracted from treebanks representative of social media language (PoSTWITA and TWITTIRÒ). Quite surprisingly, it is also the case of AlBERTo which is trained on Twitter data. 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 annota-

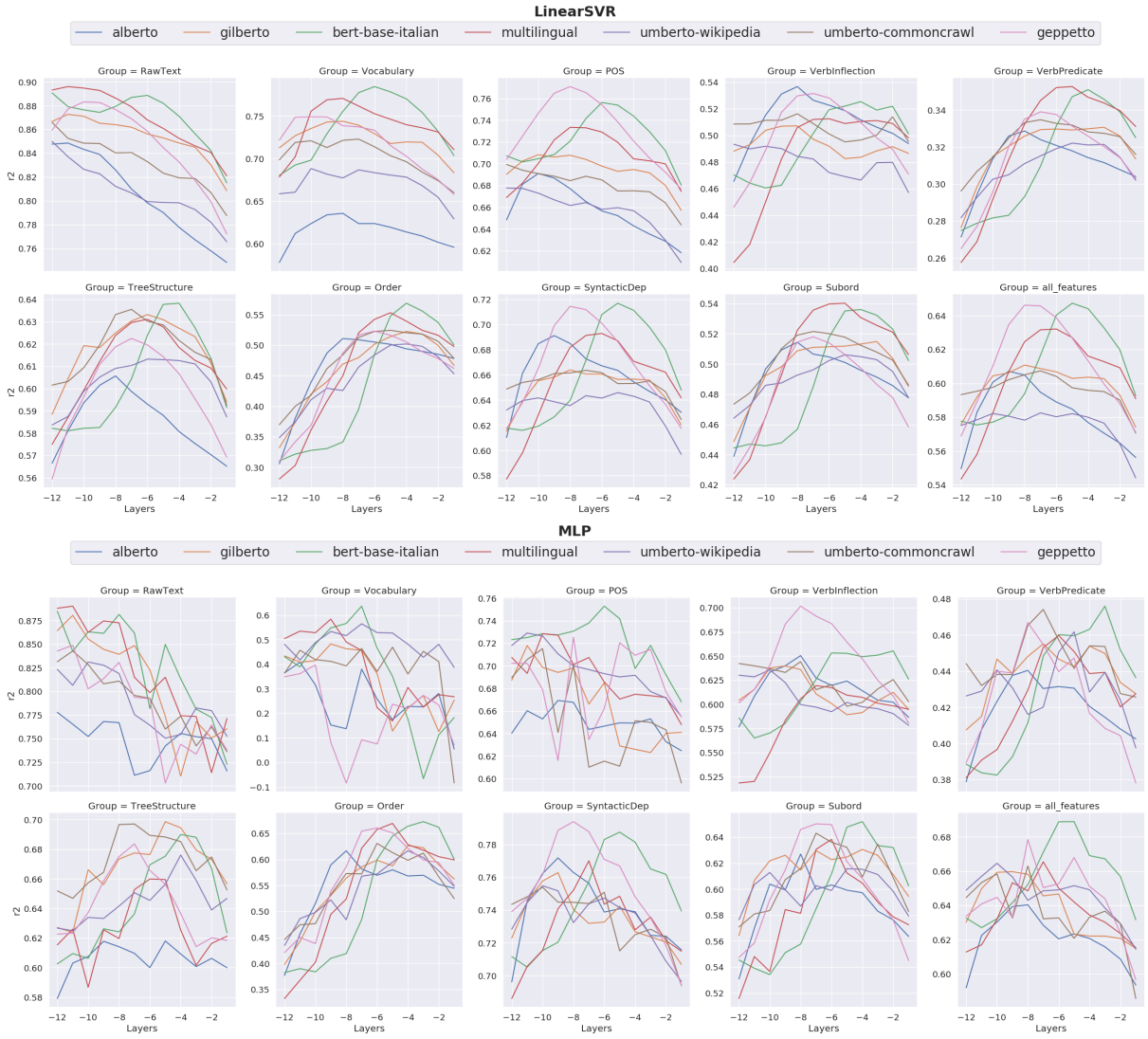

Figure 2: Average layerwise R^2 scores obtained with the LinearSVR (*top*) and the MLP (*bottom*) using the internal representations of the 7 NLMs.

tion schema. On the contrary, AlBERTo's training set is derived from Twitter's official streaming API that included all possible typologies of sentences. However, bert-base italian is slightly less affected by the non-standard linguistic peculiarities of this genre. Similarly to what is observed for the whole Italian dataset (see Figure 1), this model also reaches the highest performance in almost all different IUDT sections, except for the one containing interrogative sentences (*isdt_quest*). Interestingly, this type of sentence is 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 morphosyntactic and syntactic phenomena, hence being more difficult to encode effectively.

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 more than 80 probing features and testing our approach with two different probing models, we showed that MLP is the best model for inferring the amount of information implicitly encoded in the NLMs representations. We also observed that BERT-base-italian achieved best scores in average, but the linguistic generalization abilities of the examined transformers vary according to specific groups of linguistic phenomena and across layers. Finally, we examined how the linguistic knowledge learned by the NLMs is affected by the distinct textual varieties available in Italian tree-

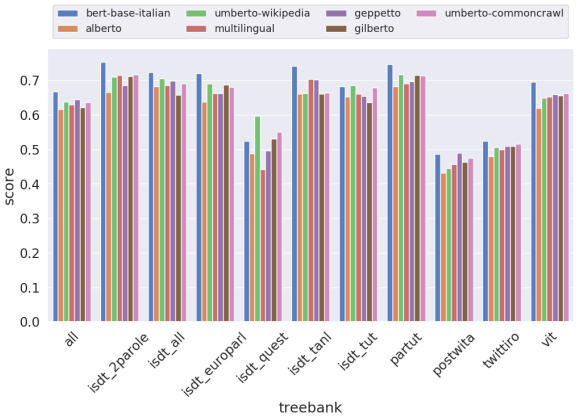

Figure 3: Average LinearSVM R^2 score considering all the UD Italian sentences (*all*) and according to the 10 treebanks previously described.

banks showing, for instance, that social media language represents a harder domain for all models.

We are currently investigating if the linguistic knowledge encoded by a NLM positively affects the resolution of downstream tasks, as already suggested by the recent work by Miaschi et al. (2020) for English. This connection, which is still rather investigated, can improve our understanding of how such models make their decisions.

References

Marco Baroni, 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.

Valerio Basile, Mirko Lai, and Manuela Sanguinetti. 2018. Long-term social media data collection at the university of turin. In *Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, pages 1–6. CEUR-WS.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

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

Terra Blevins, 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.

Cristina Bosco, Simonetta Montemagni, and Maria Simi. 2013. Converting italian treebanks: Towards an italian stanford dependency treebank. In *Proceedings of the ACL Linguistic Annotation Workshop & Interoperability with Discourse*.

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.

Alessandra Teresa Cignarella, 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)*.

Alexis Conneau, Germán 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.

Lorenzo De Mattei, Michele Cafagna, Felice Dell'Orletta, Malvina Nissim, and Marco Guerini. 2020. Geppetto carves italian into a language model. *arXiv preprint arXiv:2004.14253*.

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. *arXiv preprint arXiv:2004.06499*.

Rodolfo Delmonte, 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*.

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.

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, Hong

Kong, China, November. Association for Computational Linguistics.

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.

Ganesh Jawahar, Benoît Sagot, Djamé Seddah, Samuel Unicomb, Gerardo Iñiguez, Márton Karsai, Yannick Léo, Márton Karsai, Carlos Sarraute, Éric Fleury, et al. 2019. What does bert learn about the structure of language? In *57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy*.

Alessio Miaschi and Felice Dell'Orletta. 2020. Contextual and non-contextual word embeddings: an in-depth linguistic investigation. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 110–119, Online, July. Association for Computational Linguistics.

Alessio Miaschi, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2020. Linguistic profiling of a neural language model. *arXiv preprint arXiv:2010.01869*.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Joakim Nivre. 2015. Towards a universal grammar for natural language processing. *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 3–16.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2020. What the [mask]? making sense of language-specific bert models. *arXiv preprint arXiv:2003.02912*.

Marco Polignano, 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 *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it)*.

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

Manuela Sanguinetti, Cristina Bosco, Alberto Lavelli, Alessandro Mazzei, and Fabio Tamburini. 2018. PoSTWITA-UD: an Italian Twitter Treebank in universal dependencies. In *Proceedings of the Eleventh Language Resources and Evaluation Conference (LREC 2018)*.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. *Challenges in the Management of Large Corpora (CMLC-7) 2019*, page 9.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. 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.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019b. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

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

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

Kelly Zhang 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.