Is Neural Language Model Perplexity Related to Readability?

Alessio Miaschi[♣,⋆], Chiara Alzetta[♠,⋆], Dominique Brunato[⋆],
Felice Dell'Orletta[⋆], Giulia Venturi[⋆]
[♣]Department of Computer Science, University of Pisa
[⋆]Istituto di Linguistica Computazionale "Antonio Zampolli", ItaliaNLP Lab, Pisa
[♠]DIBRIS, University of Genoa
alessio.miaschi@phd.unipi.it, chiara.alzetta@edu.unige.it,
{name.surname}@ilc.cnr.it

Abstract

This paper explores the relationship between Neural Language Model (NLM) perplexity and sentence readability. Starting from the evidence that NLMs implicitly acquire sophisticated linguistic knowledge from a huge amount of training data, our goal is to investigate whether perplexity is affected by linguistic features used to automatically assess sentence readability and if there is a correlation between the two metrics. Our findings suggest that this correlation is actually quite weak and the two metrics are affected by different linguistic phenomena.[1]

1 Introduction and Motivation

Standard Neural Language Models (NLMs) are trained to predict the next token given a context of previous tokens. The metric commonly used for assessing the performance of a language model is perplexity, which corresponds to the inverse geometric mean of the joint probability of words $w_1, ..., w_n$ in a held-out test corpus C. While being primarily an intrinsic metric of NLM quality, perplexity has been used in a variety of scenarios, such as to classify between formal and colloquial tweets (González, 2015), to detect the boundaries between varieties belonging to the same language family (Gamallo et al., 2017) or to identify speech samples produced by subjects with cognitive and/or language diseases e.g. dementia (Cohen and Pakhomov, 2020) or Specific Language Impairment (Gabani et al., 2009). From the perspective of computational studies aimed at modeling human language processing, perplexity scores have also been shown to effectively match various human behavioural measures, such as gaze duration during reading (Demberg and Keller, 2008; Goodkind and Bicknell, 2018).

In this paper we focus on a less investigated perspective addressing the connection between perplexity and readability. Since by definition perplexity gives a good approximation of how well a model recognises an unseen piece of text as a plausible one, our intuition is that lower model perplexity should be assigned to easy-to-read sentences, while difficult-to-read ones should obtain higher perplexity. On the other hand, state-of-the-art NLMs trained on huge data have shown to implicitly learn a sophisticated knowledge of language phenomena, also with respect to complex syntactic properties of sentences (Tenney et al., 2019; Jawahar et al., 2019; Miaschi et al., 2020). This could suggest that variations in terms of linguistic complexity, especially when related to subtle morpho–syntactic and syntactic features of sentence rather than lexical ones, could not impact on model perplexity to a great extent. This assumption seems to be confirmed by the (still unpublished) results by Martinc et al. (2019) which, to our knowledge, is the only one explicitly leveraging unsupervised neural language model predictions in the context of readability assessment. According to this study, a NLM is even less perplexed by articles addressed at adults than by documents conceived for a younger readership. From a relatively different perspective focused on the ability of automatic comprehension systems to solve cloze tests, Benzahra and Yvon (2019) showed that NLMs performance is not affected by the level of text complexity.

In order to test the validity of all these hypotheses, we rely on the perplexity score given by a state-of-the-art NLM for the Italian language to several datasets representative of different textual genres containing both easy– and complex–to–read sentences: ideally, such datasets should emphasise the correlation between perplexity and readability (if present) since the corpora are explicitly designed to contain both simple and difficult examples.

Contributions We inspect whether and to which extent it is possible to find a relationship between a readability score and the perplexity of a NLM. To this aim we investigate *(i)* if the perplexity of a NLM and the readability score of a set of sentences show a significant correlation and *(ii)* whether the two metrics are equally affected by the same set of linguistic phenomena that occur in the sentence.

2 Experimental Design

According to our research questions, we devised a set of experiments to study whether NLMs perplexity reflects the level of readability of a sentence and which are the linguistic phenomena mostly involved in each metric. For this purpose, we firstly investigated whether sentence-level perplexity scores computed with one of the most prominent NLM model correlate with the scores assigned to the same sentences by a supervised readability assessment tool. Secondly, we investigated which are the linguistic features of the considered sentences that correlate in a statistically significant way with the perplexity and readability score respectively. In order to verify whether correlations hold across different typology of texts, we tested our approach on five Italian datasets.

2.1 Models

READ-IT. Automatic readability (henceforth *ARA*) was assessed using READ-IT (Dell'Orletta et al., 2011) the first readability assessment tool for Italian which combines traditional raw text features with lexical, morpho-syntactic and syntactic information extracted from automatically parsed documents. In READ-IT, analysis of readability is modelled as a binary classification task, based on Support Vector Machines using LIBSVM (Chang and Lin, 2001). Training corpora are representative of two classes of texts, i.e. difficult– vs. easy–to-read ones, both containing newspaper articles. The set of features exploited for predicting readability has been proved to capture different aspects of sentence complexity. Thus, the assigned readability score ranges between 0 (easy-to-read) and 1 (difficult-to-read) referring to the percentage probability for unseen documents or sentences to belong to the class of difficult-to- read documents. For the purposes of our work, we carried out readability assessment at sentence level, making the analysis reliable for the comparison with sentence-based perplexity of a NLM.

GePpeTto. Sentence-level perplexity scores were computed relying on GePpeTto (De Mattei et al., 2020). GePpeTto is a generative language model trained on the Italian language and built using the GPT-2 architecture (Radford et al., 2019). The model was trained on a dump of Italian Wikipedia (2.8GB) and on the itWac corpus (Baroni et al., 2009), which amounts to 11GB of web texts. The perplexity (PPL) of the model was computed as follows:

$$PPL = e^{(\frac{NLL}{N})}$$

where NNL and N correspond respectively to the negative log-likelihood and to the length of each sentence $w_{1:n} = [w_1, ..., w_n]$ in the datasets.

2.2 Corpora

In order to test the reliability of our initial hypothesis, we chose four corpora containing different typologies of texts, i.e. web pages, educational materials, narrative texts, newspaper and scientific articles. Each corpus includes a balanced amount of difficult- and easy-to-read sentence. In addition, we also considered in the analysis the Italian Universal Dependency treebank. This is meant to verify whether the connection between sentence-level readability and perplexity also holds in a well-acknowledged benchmark corpus. For each of them, we excluded from our analysis short sentences, i.e. having less than 5 tokens.

PACCSS-IT[2] (Brunato et al., 2016): we took into account 125,977 sentences belonging to PACCSS-IT, a corpus of complex-simple aligned sentences extracted from the ItWaC corpus. The resource was build using an automatic approach for acquiring large corpora of paired sentences able to intercept structural transformations (such as deletion, reordering, etc.). For example, the two following sentences represent a pair in the corpus, where a reordering operation occurs at phrase level (i.e. the subordinate clause proceeds vs. follows the main clause):

- Complex: *Ringraziandola per la sua cortese attenzione, resto in attesa di risposta.* [Lit:

[2] http://www.italianlp.it/resources/paccss-it-parallel-corpus-of-complex-simple-sentences-for-italian/

Thanking you for your kind attention, I look forward to your answer.]

- Simple: *Resto in attesa di una risposta e ringrazio vivamente per l'attenzione*. [Lit: I look forward to your answer and I thank you greatly for your attention.]

Terence and Teacher[3] (Brunato et al., 2015): two corpora of original and manually simplified texts aligned at sentence level. *Terence* contains short Italian novels for children and their manually simplified version carried out by linguists and psycholinguists targeting children with text comprehension difficulties. *Teacher* is a corpus of pairs of documents belonging to different genres (e.g. literature, handbooks) used in educational settings manually simplified by teachers. We exploited 1,644 sentences belonging to these corpora.

Multi–Genre Multi–Type Italian corpus: a collection of Italian texts representative of three traditional textual genres: Journalism, Scientific prose and Narrative. Each genre has been internally subdivided into two sub-corpora representative of an easy- vs difficult-to-read variety, which was defined according to the intended target audience for a given genre. The journalistic prose corpus includes articles automatically downloaded from the online versions of two general-purpose newspapers[4], while the "easy" sub-corpus contains articles from two easy-to-read newspapers[5] addressed to adults with low literacy skills or mild intellectual disabilities. The scientific prose collection consists of scholarly publications on linguistics and computational linguistics and Wikipedia pages downloaded from the portal "Linguistics", representative of the complex and easy variety respectively. For the narrative genre, we included long novels written by novelists of the last century and contemporary writers in the corpora of complex variety, while for the easy variety we collected short novels for children. The complete corpus contains 56,685 sentences.

Italian Universal Dependency Treebank: it includes different sections of the Italian Universal Dependency Treebank (IUDT), version 2.5 (Zeman et al., 2019). In particular, we considered two groups: a first one containing the whole Italian Stanford Dependency Treebank (ISDT)[6] (Bosco et al., 2013), the Italian version of the multilingual Turin University Parallel Treebank (Sanguinetti and Bosco, 2015) and the Venice Italian Treebank (Delmonte et al., 2007) (24,998 sentences), all containing a mix of textual genres; and a second one including two collections of texts representative of social media language, i.e. generic tweets and tweets labelled for irony (PosTWITA[7] and TWITTIRO[8]) (Sanguinetti et al., 2018; Cignarella et al., 2019) (3,660 sentences in total).

3 Sentence Perplexity and Readability

Our analysis starts from a comparison between the average perplexity and readability scores obtained for each sentence of the five considered datasets. As shown in Table 1, readability values (column *ARA*) are quite homogeneous across the datasets, with low standard deviation values. On the contrary, the range of perplexity scores is wider (column *PPL*), going from an average score of 3,905.83 of PACCSS-IT to 436.75 of the IUDT miscellaneous portion (Italian UD). These differences seem to provide a first evidence that perplexity and readability are not correlate to each other.

This intuition has been proved computing the Spearman's rank correlation coefficient between the perplexity and readability scores for each dataset. Results are reported in Table 2, column *PPL-ARA*. As it can be seen, all correlation rates are significant, except for the result obtained on the Terence and Teacher corpus, possibly due to the fact that the size of the corpus is too small to allow a significant comparison. Contrary to our expectations, no correlation was detected between the two metrics for all corpora, suggesting that perplexity and and readability are independent from each other.

To further investigate the reasons behind these scores and to deepen the analysis about the relationship between the two metrics, we investigated whether they capture the same (or similar) linguistic properties of the sentences. To this aim, we tested the presence and strength of the correlation between each of the two metrics and a set of 176 linguistic features, which have been shown to capture properties of sentence complex-

[3] http://www.italianlp.it/resources/terence-and-teacher/

[4] www.repubblica.it and http://www.ilgiornale.it/

[5] www.dueparole.it and http://www.informazionefacile.it/

[6] https://github.com/UniversalDependencies/UD_Italian-ISDT

[7] https://github.com/UniversalDependencies/UD_Italian-PoSTWITA

[8] https://universaldependencies.org/treebanks/it_twittiro

Dataset	**PPL**	**ARA**
PACCSS-IT	3,905.83 (± 21,306.07)	0.55 (± 0.24)
Terence-Teacher	790.85 (± 5,002.62)	0.46 (± 0.27)
Multi-Genre Multi-Type	570.85 (± 4,820.12)	0.58 (± 0.31)
Italian-UD	436.75 (± 3,633.64)	0.61 (± 0.30)
Twitter-UD	986.28 (± 2,479.64)	0.59 (± 0.30)

Table 1: Perplexity (*PPL*) and Readability (*ARA*) mean and standard deviation values for the 5 datasets.

Dataset	**PPL-ARA**	**Feats**
PACCSS-IT	-0.031*	0.169*
Terence-Teacher	0.014	0.149
Multi-Genre Multi-Type	0.026*	0.184*
Italian-UD	-0.054*	0.332*
Twitter-UD	-0.038*	-0.037

Table 2: Spearman's correlation coefficients between sentence-level perplexity and readability scores (*PPL-ARA*) and between rankings of linguistic features (*Feats*). Statistically significant correlations ($p < 0.05$) are marked with *.

ity (Brunato et al., 2018). In particular, this analysis is based on the set of features described in Brunato et al. (2020), which are acquired from raw, morpho-syntactic and syntactic levels of annotation. They range from basic information on the average sentence and word length, to lexical information about the internal composition of the vocabulary of the text (e.g. the distribution of lemmas belonging to the *Basic Italian Vocabulary* (De Mauro, 2000)). They also include morpho–syntactic information (e.g. POS distribution and of inflectional properties of verbs) and more complex aspects of sentence structure derived from syntactic annotation and modeling global and local properties of parsed tree structure, e.g. the relative order of subjects and objects with respect to the verb, the use of subordination. In order to extract these features, the considered corpora were morpho-syntactically annotated and dependency parsed by the UDPipe pipeline (Straka et al., 2016), with the exception of the IUDT corpus.

Column *Feats* of Table 2 illustrates the results of this analysis: we report the Spearman's correlation coefficients between the two rankings of linguistic features, each ordered by strength of correlation between feature value and perplexity score and readability score respectively. Once again we observe rather weak correlation values, with the only exception of Italian-UD which is the only one reporting a medium correlation (.332). Overall, these results corroborate our previous findings that the two metrics are not particularly related with each other, and they further suggest that the linguistic phenomena affecting the perplexity of NLM and the readability level of a sentence are very different. Consider for example the two following sentences:

(1) *Il furto è avvenuto giovedì notte.*
The theft has taken place Thursday night.

(2) *Il comitato di bioetica: no all'eutanasia.*
The bioethics committee: no to euthanasia.

While (1) is very easy-to-read, with a readability score of 0.25, but it has a quite high perplexity score, i.e. 40,737.81, (2) is quite difficult-to-read (ARA=1) but is has a very low perplexity score (PPL=11.24).

4 In-Depth Linguistic Investigation

To better explore the motivation behind these results, we performed an in-depth investigation aimed at understating the relationship between our set of linguistic features and the two metrics taken into consideration. Since we noticed that for all datasets a higher number of features correlates with ARA than with PPL, we selected those that are significantly correlated with the two metrics. The number of shared features varies for each dataset, depending on their size. For example, for the two smallest ones, i.e. Terence and Teacher and the UD Twitter Treebank, we could only consider 34.65% (61) and 44.88% (79) of the whole set of features respectively, while for the larger corpora the sub-set is wider: 81.81% (144) in PACCSS-IT, 78.97% (139) for Multi-Genre Multi-Type and 84.65% (149) for the IUD Treebank.

Table 3 shows the top ten features for each dataset, i.e. those that obtained the strongest correlation with both PPL and ARA. As expected, correlations are generally stronger between linguistic features and readability scores, although they are lower than expected. This could be due to the fact that, even if the READ–IT classifier is trained with a similar set of features, the non-linear feature space makes it difficult to identify clear correlations with individual features. Similarly, our set of features seem to play only a marginal role on perplexity. However, this is not the case of the PACCSS-IT corpus, for which the set of considered linguistic features have an higher correlation with PPL. This can be possibly related to the partial overlap between the GePpeTto training data and the PACCSS-IT sentences, since the latter is drawn from the ItWac corpus which is included in the GePpeTto's training.

Inspecting these results, we can also observe that correlations between features and PPL seem to be more affected by genre–specific characteristics. This is particularly clear if we consider the Italian UD Twitter treebank, for which among the top ten most correlated features we find some of them characterising social media language, e.g. symbols (*upos-xpos_dist_SYM*) or the vocative relation, which marks a dialogue participant addressed in a text along with the specification, specifically used for Twitter @-mentions (*dep_dist_vocative:mention*).

5 Conclusion

The paper presented a study aimed at investigating the relationship between two metrics computed at sentence-level, i.e. perplexity of a state-of-the-art NLM for the Italian language and readability score automatically assigned to a sentence by a supervised classifier. We carried out our analysis considering several datasets differing at the level of textual genre and language variety. Specifically, we observed that comparing the rankings obtained using the two metrics we cannot find any significant correlation, either between the scores of the two metrics or with respect to the set of linguistic features that mostly impact their values. Further investigation within this line of research will explore whether we can draw the same observations when a different NLM is exploited to compute sentence perplexity.

PACCSS-IT			
PPL		**ARA**	
Feats	*Corr*	*Feats*	*Corr*
aux_num_pers_dist_Sing+3	0,53	xpos_dist_FF	0,34
dep_dist_cop	0,51	dep_dist_punct	0,32
avg_max_depth	0,50	upos_dist_PUNCT	0,32
upos_dist_ADP	0,50	ttr_form	0,29
xpos_dist_E	0,50	aux_mood_dist_Cnd	0,25
dep_dist_case	0,49	upos_dist_DET	0,25
n_tokens	0,48	dep_dist_det	0,25
dep_dist_root	0,48	ttr_lemma	0,22
xpos_dist_FS	0,48	upos_dist_NOUN	0,21
Terence and Teacher			
PPL		**ARA**	
Feats	*Corr*	*Feats*	*Corr*
xpos_dist_B	0,25	dep_dist_det	-0,39
verbs_num_pers_dist_Sing+3	0,23	upos_dist_DET	-0,38
lexical_density	0,22	upos_dist_NOUN	-0,37
dep_dist_advmod	0,21	xpos_dist_S	-0,37
upos_dist_ADV	0,21	xpos_dist_RD	-0,29
verbs_num_pers_dist_Plur+3	-0,16	upos_dist_ADV	0,27
xpos_dist_V	0,16	dep_dist_advmod	0,25
avg_token_per_clause	-0,16	xpos_dist_FF	0,25
upos_dist_VERB	0,14	avg_sub_chain_len	0,24
Multi-Genre Multi-Type			
PPL		**ARA**	
Feats	*Corr*	*Feats*	*Corr*
n_tokens	-0,19	principal_prop_dist	-0,42
dep_dist_root	0,19	ttr_form	-0,34
dep_dist_advmod	0,19	xpos_dist_FF	0,34
upos_dist_ADV	0,18	dep_dist_det	-0,33
n_prepositional_chains	-0,18	upos_dist_DET	-0,33
xpos_dist_B	0,18	upos_dist_PUNCT	0,33
upos_dist_ADP	-0,17	dep_dist_punct	0,33
xpos_dist_E	-0,17	xpos_dist_FB	0,31
ttr_lemma	0,16	sub_prop_dist	0,27
Italian UD Treebank			
PPL		**ARA**	
Feats	*Corr*	*Feats*	*Corr*
n_tokens	-0,27	principal_prop_dist	-0,53
dep_dist_root	0,27	sub_proposition_dist	0,40
n_prepositional_chains	-0,26	n_tokens	0,39
avg_max_depth	-0,24	dep_dist_root	-0,39
upos_dist_ADP	-0,24	ttr_form	-0,37
ttr_lemma	0,23	avg_max_depth	0,36
max_links_len	-0,23	avg_links_len	0,35
avg_max_links_len	-0,23	max_links_len	0,34
xpos_dist_E	-0,22	avg_max_links_len	0,34
Italian UD Twitter Treebank			
PPL		**ARA**	
Feats	*Corr*	*Feats*	*Corr*
upos_dist_SYM	0,38	upos_dist_PUNCT	0,30
avg_max_depth	-0,28	dep_dist_punct	0,30
xpos_dist_SYM	0,28	dep_dist_det	-0,29
in_dict	-0,24	upos_dist_DET	-0,29
dep_dist_vocative:mention	0,23	verbal_root_perc	-0,27
in_dict_types	-0,22	xpos_dist_RD	-0,27
ttr_lemma	0,21	avg_token_per_clause	-0,27
in_FO	-0,21	subj_pre	-0,27
verbal_head_per_sent	-0,19	obj_post	-0,24

Table 3: Top 10 features along with their correlation scores between perplexity and readability.

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.

Marc Benzahra and François Yvon. 2019. Measuring text readability with machine comprehension: a pilot study. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 412–422, Florence, Italy, August. Association for Computational Linguistics.

C. Bosco, S. Montemagni, and M. Simi. 2013. Converting italian treebanks: Towards an italian stanford dependency treebank. In *Proceedings of the ACL*

Linguistic Annotation Workshop & Interoperability with Discourse, Sofia, Bulgaria, August.

Dominique Brunato, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2015. Design and annotation of the first italian corpus for text simplification. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 31–41.

Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, and Giulia Venturi. 2016. PaCCSS-IT: A parallel corpus of complex-simple sentences for automatic text simplification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 351–361, Austin, Texas, November. Association for Computational Linguistics.

Dominique Brunato, Lorenzo De Mattei, Felice Dell'Orletta, Benedetta Iavarone, and Giulia Venturi. 2018. Is this sentence difficult? do you agree? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2690–2699, Brussels, Belgium, October-November. Association for Computational Linguistics.

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 7145–7151, Marseille, France, May. European Language Resources Association.

Chih-Chung Chang and Chih-Jen Lin. 2001. LIBSVM: a library for support vector machines.

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)*.

Trevor Cohen and Serguei Pakhomov. 2020. A tale of two perplexities: Sensitivity of neural language models to lexical retrieval deficits in dementia of the Alzheimer's type. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1946–1957, Online, July. Association for Computational Linguistics.

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*.

Tullio De Mauro. 2000. *Il dizionario della lingua italiana*, volume 1. Paravia.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. READ–IT: Assessing readability of Italian texts with a view to text simplification. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83, Edinburgh, Scotland, UK, July. Association for Computational Linguistics.

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*.

V. Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109:193–210.

Keyur Gabani, Melissa Sherman, Thamar Solorio, Yang Liu, Lisa Bedore, and Elizabeth Peña. 2009. A corpus-based approach for the prediction of language impairment in monolingual English and Spanish-English bilingual children. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 46–55, Boulder, Colorado, June. Association for Computational Linguistics.

Pablo Gamallo, Jose Ramom Pichel, and Iñaki Alegria. 2017. A perplexity-based method for similar languages discrimination. In *VarDial2017 workshop at EACL 2017. Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 109–114,Valencia, Spain, April 3, 2017. c©2017 Association for Computational Linguistics (http://web.science.mq.edu.au/ smalmasi/vardial4/index.html)*.

M. González. 2015. An analysis of twitter corpora and the differences between formal and colloquial tweets. In *TweetMT@SEPLN*.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In Asad B. Sayeed, Cassandra Jacobs, Tal Linzen, and Marten Van Schijndel, editors, *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics, CMCL 2018, Salt Lake City, Utah, USA, January 7, 2018*, pages 10–18. Association for Computational Linguistics.

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*.

Matej Martinc, Senja Pollak, and Marko RobnikSikonja. 2019. Supervised and unsupervised neural approaches to text readability. *Computing Research Repository, arXiv:1503.06733. Version 2*.

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

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

Manuela Sanguinetti and Cristina Bosco. 2015. Part-TUT: 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)*.

M. Straka, J. Hajic, and J. Strakova. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*.

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

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)*.