Exploring the Relationship between Sparse Autoencoder Activations and Sentence Complexity

Ejdis Gjinika[1,*], Luca Putelli[1,*], Ivan Serina[1] and Alfonso Emilio Gerevini[1]

[1]*Università degli Studi di Brescia, Via Branze 38, Brescia, Italy*

Abstract

Sparse Autoencoders (SAEs) have recently emerged as a promising approach for interpreting the internal word representations of Neural Language Models (NLMs). Existing approaches focus on analysing the sparse activations produced by SAEs, verifying whether they encode interpretable concepts. However, how SAEs activations reflect different texts and their linguistic complexity remains unclear. We investigate this phenomenon considering three datasets characterized by different aspects of complexity and we analyse it across the layers of two small NLMs. We leverage a metric that quantifies the Number of Effective Features (NEF) that contribute to SAE activations for a given complexity level. Comparing the different results, we found that NEF values are correlated with sentence complexity in the middle-to-last layers while in early layer this trend is reversed.

Keywords

Representation analysis, Sentence Complexity, Sparse Autoencoders, Neural Language Models

1. Introduction

Understanding how Neural Language Models (NLMs) encode linguistic properties in their internal representations remains an open question in mechanistic interpretability. In this scope, Sparse Autoencoders (SAEs) have recently emerged as a promising tool for decomposing neural representations into interpretable features [2, 3]. Given the vectorial representation of a portion of text provided by a NLM, the key idea of this technique is to feed the vector in a SAE, which (as in traditional autoencoders) has to reproduce it in output. However, during this operation the SAE has to project it into a larger space (using a feed-forward neural network layer), called the sparse feature space. Ideally, this space is composed by many numeric features (i.e. vector components), and only a few are activated for each input. The main claim is that such features represent interpretable concepts present in the original text [4]. With this technique, several properties, such as synonyms [5] and verb tenses [6] have been shown to be encoded by language model representations.

However, there is very little information about how SAEs behave with different input and contexts. How do these models react with sentences of different length? Are these features mainly linguistic? Are they related to the text topic? In this work, we propose a preliminary study on these subjects, and in particular we investigate the behaviour of SAEs in the context of linguistic complexity, analysing whether more complex sentences activate a larger number of SAE features. Linguistic complexity has no single agreed definition, spanning lexical, syntactic, and semantic aspects of language [7, 8]. Ideally, a complex sentence is composed either of difficult terms or has subordinate clauses. We expect that such complexity should be reflected in the internal representations of language models, and such sentences may require the model to encode a larger variety of linguistic phenomena, including richer semantic relations and more complex syntactic structures. Therefore, we hypothesize that complex sentences should involve a more diverse set of SAE features.

To investigate this hypothesis, we study whether differences in linguistic complexity are reflected in sparse representations across model layers. We quantify these differences through the Number of

NL4AI 2026: Ninth Workshop on Natural Language for Artificial Intelligence, October 6-9th, 2026, Perugia, Italy [1]

*Corresponding author.

ejdis.gjinika@unibs.it (E. Gjinika); luca.putelli@unibs.it (L. Putelli); ivan.serina@unibs.it (I. Serina); alfonso.gerevini@unibs.it (A. E. Gerevini)

0009-0006-9817-5846 (E. Gjinika); 0009-0008-5055-6812 (L. Putelli); 0000-0002-7785-9492 (I. Serina); 0000-0001-9008-6386 (A. E. Gerevini)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).



published 2026-11-05

CEUR-WS.org/Vol-4315/paper17.pdf

Effective Features (NEF), which estimates how many SAE features effectively contribute to a sparse representation [9]. To this end, we consider and adapt three datasets designed to capture different aspects of linguistic complexity. These datasets are annotated with different objectives. The CEFR dataset [10] reflects the overall complexity of the sentence in terms of human proficiency, the Simplification [11] dataset distinguish between original complex sentences and their simplified versions, while Lexical Complexity [12] dataset focuses on the presence of one or more "difficult" terms in a sentence. In our analysis, we exploit two pre-trained Pythia NLMs [13] with 70M and 160M parameters, using publicly available SAEs trained on their corresponding model activations. Our results show that linguistic complexity is reflected in sparse representations, with more complex sentences generally exhibiting higher NEF values in intermediate and later layers, with an opposite trend in the early layers.

2. Related work

Interpretability methods aim to uncover how textual representations capture linguistic and semantic knowledge. Among them, probing methods assess whether specific concepts can be recovered from vectorial representations [14], and have been widely applied to study syntactic [15, 16] and semantic properties [17, 18].

More recently, Sparse Autoencoders (SAEs) have emerged as a powerful method of interpretability, in particular, in extracting features from representations produced by a Neural Language Model (NLM) [2]. One of the key contributions in this area introduced the use of a Top-k activation mechanism in SAEs, retaining only the k largest activations for each input [19]. Building on this architecture, several works have released open suites of pre-trained SAEs for production-scale models, enabling systematic interpretability research without the substantial computational cost of training dictionaries from scratch [20, 21]. Generally, SAEs are used to identify interpretable features associated with concepts or patterns [3], or to decompose the dense activations of NLMs into sparse representations for understanding their internal mechanisms [2]. With respect to these works, we do not focus on the interpretation of individual features, but rather analyse the organization of sparse representations and how it changes across different levels of linguistic complexity.

Measuring text complexity is a challenging task, as the perceived difficulty of a text depends on subjective judgments and varies across annotators [22, 23]. Therefore, defining whether a text is easy or difficult to understand is not a straightforward process. A formal perspective on natural language complexity is provided by [24], who study complexity levels from the point of view of computational linguistics and formal language theory. The work in [25] defines sentence complexity in terms of surface-level linguistic properties, such as sentence length, vocabulary choice, and syntactic structure.

3. Datasets Construction

To perform the analyses, we considered and adapted three datasets in which sentences are labeled according to their complexity level. Table 1 summarizes the main characteristics of the three datasets.

CEFR Dataset The CEFR dataset is based on the *cefr_sp_en* dataset that is part of the *UniversalCEFR* dataset [26]. The Common European Framework of Reference for Languages (CEFR) is an internationally recognized standard for describing language proficiency. It defines six proficiency levels, ranging from A1 for beginner learners to C2 for proficient users. These levels are widely used to assess the complexity of language production and comprehension. The CEFR dataset contains about $10,000$ sentences with their complexity level annotated by two experts. For instance, the sentence *"Foxes often live in forests."* is categorized as A1, whereas *"Lithium oxide is produced by thermal dehydration of lithium hydroxide."* is categorized as C2. For our experiments, we selected 120 sentences for each of the 6 complexity levels.

Simplification Dataset The Simplification dataset is derived from the *Asset* dataset published in [11], which contains more than $2,000$ complex sentences, with ten different simplified versions for

Table 1
Word-length statistics (mean, standard deviation, and maximum number of words) per sentence, computed separately for each label of every dataset. CEFR sentences are grouped by proficiency level (A1–C2), while Simplification and Lexical Complexity sentences are grouped by Simple/Complex label.

	CEFR						Simplification		Lexical Complexity	
Statistics	A1	A2	B1	B2	C1	C2	Simple	Complex	Simple	Complex
Mean	8.40	9.91	12.62	16.68	18.35	19.29	16.70	19.28	23.25	27.77
Std	3.13	3.71	4.10	5.72	5.94	6.06	6.84	7.77	12.35	14.78
Max	18	23	24	30	30	33	41	45	67	95

each of them written by human annotators. Our dataset has been built by randomly sampling 1180 complex sentences, such as "*The West Coast blues is a type of blues music characterized by jazz and jump blues influences, strong piano-dominated sounds and jazzy guitar solos, which originated from Texas blues players relocated to California in the 1940s.*", and 1180 simple sentences, such as "*The town lies between Baden and Zurich.*".

Lexical Complexity Dataset Our Lexical Complexity Dataset originates from the *CompLex* dataset introduced in [12]. In CompLex, each sentence contains one or more annotated words, each assigned by a human annotator with a complexity score between 0 and 1, where higher values indicate greater complexity. We process this dataset by averaging the complexity scores of the annotated words within each sentence. We then divide the resulting instances into two groups: simple sentences, with an average score below 0.25, and complex sentences with an average score above 0.75. For example, "*He brought them forth with silver and gold.*" has a complexity score of 0.02, whereas "*Most of the atypical splice forms we observe appear to encode non-functional transcripts, containing frameshifts or lacking a start codon or other functional residues conserved throughout the olfactory receptor family.*" has a complexity score of 0.83. This datasets consists of 110 sentences of each label.

4. Analysing Complexity of Sparse Representations

In this work, our goal is to investigate how the Number of Effective Features (NEF) varies across sentences with different levels of complexity. In this context, a feature corresponds to a latent dimension learned by the SAE. Intuitively, the NEF measure the number of features (and therefore, as claimed by [2], interpretable concepts) that can be extracted from the sentences in a dataset. We perform this analysis across all layers of two NLMs models, and on the three datasets introduced in Section 3.

4.1. Analysis Framework

The proposed framework is shown in Figure 1 consists of three main stages: (i) extracting representations from the NLMs, (ii) projecting these representations into a sparse feature space using pretrained SAEs, and (iii) analysing the resulting feature activations.

Given a dataset composed by S sentences, into which the $S_\ell \subset S$ are the sentences associated with the label ℓ, we feed each sentence $s \in S_\ell$ into a NLM, which generates a representation for each token of the sentence. Since our aim is to examine the entire sentence, we compute the mean of all token representations to obtain a single representation for each sentence. The extraction process is conducted over all the layers of the model except the embedding layer. Once extracted, the representations are fed into the SAE which produces the sparse representations, called z_s. We extract the sparse representations for each sentence and for each layer using the SAE trained on that specific layer.

To measure the NEF that represent a given label, we compute it as described by [9]. Considering a SAE in which the internal, sparse layer, is made by D components, for each component $i \in D$, we

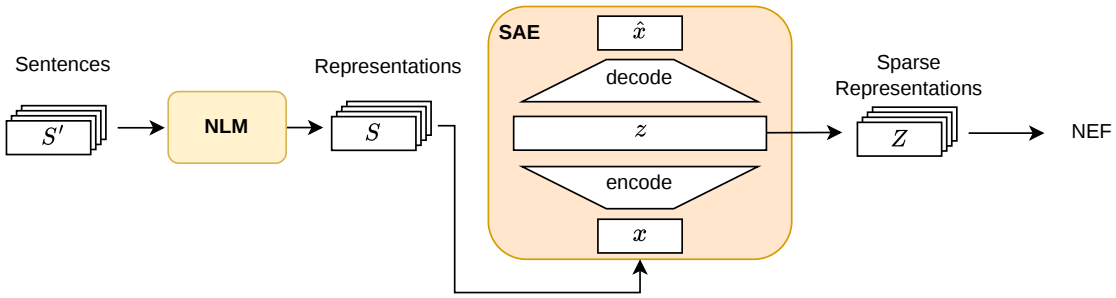


Figure 1: Overview of the proposed method. Sentences are first encoded into dense representations by the NLM; these are then mapped by the SAE into sparse representations, from which the NEF is computed by considering the distribution of activation magnitudes across features.

calculate a probability function p_i defined as:

$$p_i = \frac{\sum_{s=1}^{S_\ell} |z_{i,s}|}{\sum_{j=1}^{D} \sum_{s=1}^{S_\ell} |z_{j,s}|}$$

in which $z_{i,s}$ denotes the activation of sparse feature i on sample s. Next, we calculate the Shannon entropy of the probability distribution p_i, and the number of effective features (NEF) for the label ℓ is computed as the exponential of the Shannon entropy :

$$H_\ell(p) = -\sum_{i=1}^{D} p_i \log p_i, \quad \text{NEF}_\ell = e^{H_\ell(p)}$$

Lower values of NEF_ℓ indicate that representing the sentences belonging to a label ℓ require a small number of sparse features. Conversely, higher values of NEF_ℓ suggest that the representation of a label require representing more concepts, and therefore more sparse features.

5. Experimental Evaluation

In our analysis, we considered the 70M and 160M models of the Pythia suite [13], together with their corresponding pretrained SAEs. In this context, the pretrained SAEs are trained following the setting introduced by [19]. This implies that SAEs are trained with Top-k activation, so all sparse representations have at most k active features (i.e. with a value greater than zero). The models produce hidden representations with dimensions of 512 and 768 for the 70M and 160M models, respectively. The corresponding SAEs have a latent dimension of 32,768, with a k value of 16 and 32 for the 70M and 160M models, respectively. For both these models, we investigate how NEF varies across the CEFR, Simplification and Lexical Complexity datasets.

5.1. Results

The results are available in Figure 2. Across all three datasets, NEF values consistently varies with sentence complexity. In the following, we analyse the NEF values, and therefore the SAEs behaviour, dividing the models in two different regions: the early layers and the middle-to-late layers. We categorize as early layers the first two layers for the 70M and the first three layers for the 160M, whereas we define middle-to-late layers as all the subsequent ones. Finally, we investigate whether the observed trends are consistent across model sizes. For both models, the reconstruction error (measured in terms of Fraction of Variance Unexplained, FVU) is low across most layers (ranging from 0.2 and 0.7), except for a rise at the final layer (FVU nearly 1.5 for 160M and greater than 4.5 for the 70M).

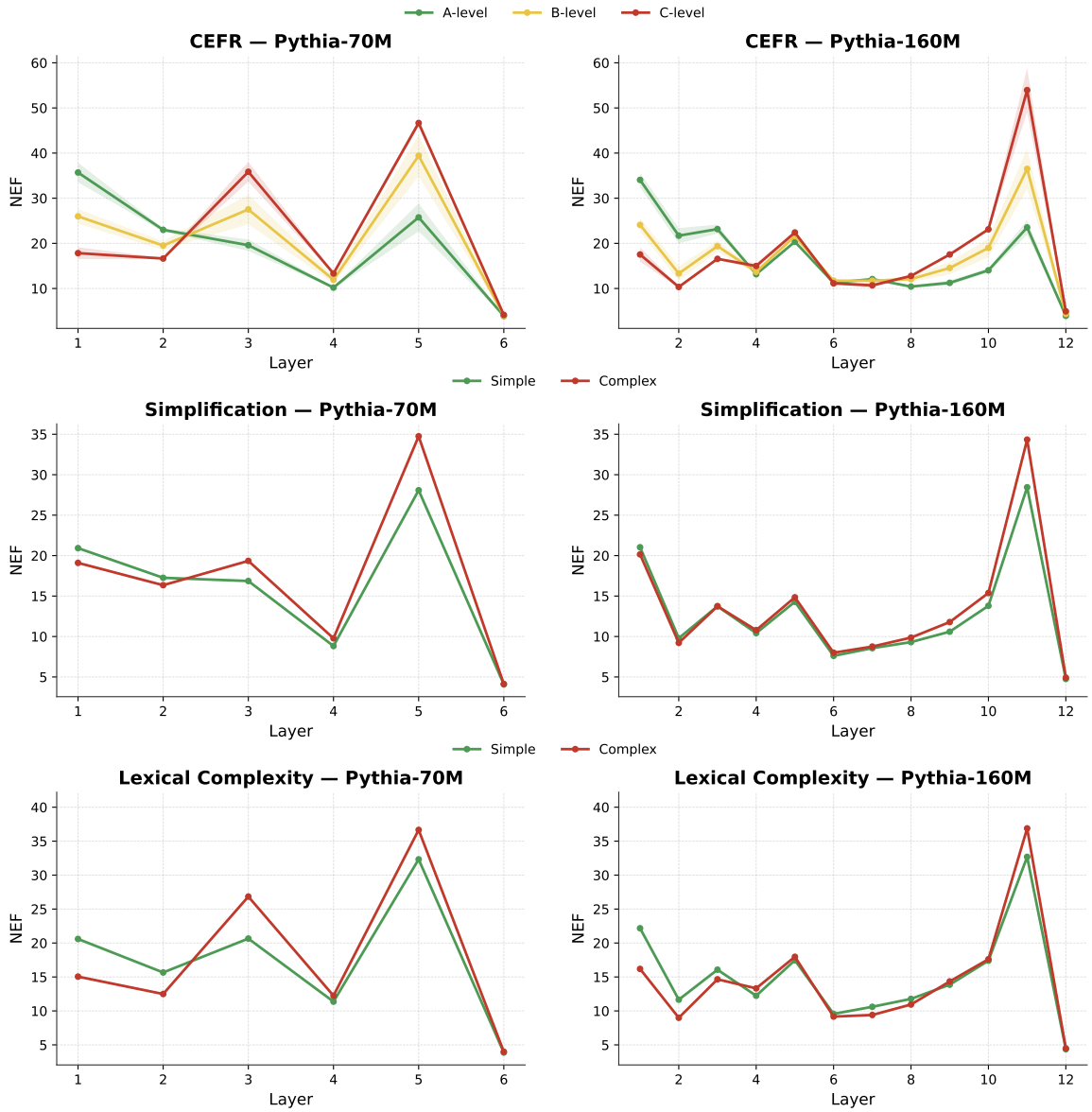


Figure 2: Values of NEF (on the y-axis) across the three datasets (CEFR, Simplification and Lexical Complexity) and layers (on the x-axis) of both analysed models. For the CEFR dataset, we aggregate the six single complexity levels and report the mean and standard deviation of the A-level, B-level and C-level.

Early layers In early layers, simpler sentences have higher NEF values than the complex ones. This behaviour is consistent in both models and is particularly evident in CEFR, considering the first two layers of the 70M model and the first three layers of the 160M model. For instance, A1-level sentences exhibit a NEF of approximately 35.7, compared to 17.8 for C2-level sentences in the 70M model. In the $160M$ we have 34.0 for A1 and 17.5 for C2. The Lexical Complexity dataset shows a similar behaviour, with a difference (in favour of the simpler sentences) of 5.5 for the first layer and 3.2 for the second layer of the $70M$. Similarly, the $160M$ obtains differences of 5.9, 2.6 and 1.4 for the first three layers. This behaviour is less pronounced in the Simplification dataset, where the difference in NEF values between simple and complex sentences is below 1.8 for the 70M model and 1.0 for the 160M model.

One possible explanation is that the early layers concentrate on basic syntactic concepts that are highly present in simple sentences, resulting in a more distributed representation of simple structures rather than complex linguistic phenomena, coherently with what has been observed in [27, 28, 29]. In fact, please note that the highest average NEF values in the early layers, for both models, are obtained by the A-level sentences of the CEFR dataset, which are very basic and have (as it can be seen in Table 1)

on average less than 10 words. Slightly more complex sentences (B-level, or the simple sentences contained in the other two datasets), consistently have a higher number of words and they obtain lower values of NEF, suggesting that they may contain more intricate concepts. However, this explanation should be regarded as an hypothesis, and further investigation is needed.

Middle-to-late layers We observe that in the middle-to-late layers, the NEF is correlated with the complexity. This behaviour is consistent across both models and is particularly evident in the CEFR and Simplification datasets. For these two datasets, we can note this pattern starting from layers 3 and 8 for 70M and 160M, respectively. The largest difference in NEF is observed in the penultimate layer of both models. Considering the $70M$ model on the CEFR dataset, the NEF of the A-level reaches 25.7, whereas the C-level reaches 46.6 for the 70M model. Similarly, the 160M model exhibits an increase from 23.5 to 53.9. The same trend is observed on the Simplification dataset, although the increase is less pronounced. The NEF consistently grows with complexity, with similar increments in the two models. Specifically, the 70M model increases from 28.0, obtained by simple sentences, to 34.7 obtained by the complex ones, while the 160M model increases from 28.4 to 34.3. This indicates that sets of complex sentences activate a more diverse set of feature, with respect than sets of simpler sentences. In the Lexical Complexity dataset this pattern is visible in the third and in the penultimate layer of the 70M, with a difference between simple and complex sentences of 6.2 and 4.3, respectively. In the $160M$, simple and complex sentences obtain basically the same NEF values across all late layers, with only a small difference in the penultimate one (4.2). This peculiar behaviour might be due to the dataset annotation methodology, which involves averaging the complexity of words, rather than evaluating the overall sentence complexity. A notable exception across all models and datasets, is that the last layer always obtain almost identical NEF values (between 4 and 5). This phenomenon is consistent with previous studies showing that the last layer is only responsible for language modeling, rather than containing linguistic or semantic concepts [30, 18, 16].

Finally, another observation is that despite a substantial different of our two models in terms of their size, both of them obtain comparable NEF values, especially for the Simplification and Lexical Complexity datasets. Even the peak NEF values at the penultimate layer differ by less than 0.4 for both simple and complex labels. This might indicate that both models have sufficient capacity to represent the sentences in these datasets, whereas the difference in CEFR (in which 160M can reach higher values with respect to 70M), might indicate that 70M is approaching its representational limits for the most complex sentences. However, this is an hypothesis which needs to be further investigated.

6. Conclusions and Future Work

In this work, we analysed whether sentence complexity is reflected in sparse representations across three datasets and two small NLMs in terms of Number of Effective Features. Our experiments reveal three main findings. First, sentence complexity is associated with higher NEF values in middle-to-late layers. Second, this relationship is reversed in the early layers, into which higher NEF values are obtained by simple sentences. Interestingly, Simplification and Lexical Complexity datasets are made by sentences with very similar length. Therefore, NEF might be not influenced by sentence length or other superficial characteristics, but more by the actual complexity of the sentence. As future work, we will address sentence complexity from a more general perspective, including more datasets and models. Moreover, we aim to understand why the behaviour of sparse representations changes from early to the later layers, and its relation with sentence length.

Declaration on Generative AI

During the preparation of this work, the authors have employed ChatGPT for grammar and spelling check. All content was reviewed by the authors, who take full responsibility for the publication's content.

References

[1] A. Bondielli, G. Bonetta, E. Leonardelli, I. Siragusa, Preface to the Ninth Workshop on Natural Language for Artificial Intelligence (NL4AI), in: Proceedings of the Ninth Workshop on Natural Language for Artificial Intelligence (NL4AI 2026) co-located with 24th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2026), 2026.

[2] H. Cunningham, A. Ewart, L. Riggs, R. Huben, L. Sharkey, Sparse autoencoders find highly interpretable features in language models, 2023.

[3] A. Templeton, T. Conerly, J. Marcus, J. Lindsey, T. Bricken, B. Chen, A. Pearce, C. Citro, E. Ameisen, A. Jones, H. Cunningham, N. L. Turner, C. McDougall, M. MacDiarmid, C. D. Freeman, T. R. Sumers, E. Rees, J. Batson, A. Jermyn, S. Carter, C. Olah, T. Henighan, Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet, Transformer Circuits Thread (2024).

[4] N. Elhage, T. Hume, C. Olsson, N. Schiefer, T. Henighan, S. Kravec, Z. Hatfield-Dodds, R. Lasenby, D. Drain, C. Chen, R. Grosse, S. McCandlish, J. Kaplan, D. Amodei, M. Wattenberg, C. Olah, Toy models of superposition, 2022.

[5] A. Diera, A. Scherp, Do language models encode semantic relations? probing and sparse feature analysis, 2026.

[6] Y. Jing, Z. Yao, H. Guo, L. Ran, X. Wang, L. Hou, J. Li, LinguaLens: Towards interpreting linguistic mechanisms of large language models via sparse auto-encoder, in: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Suzhou, China, 2025, pp. 28232–28251.

[7] S. Pollard, A. W. Biermann, A measure of semantic complexity for natural language systems, in: NAACL-ANLP 2000 Workshop: Syntactic and Semantic Complexity in Natural Language Processing Systems, 2000.

[8] K. North, M. Zampieri, M. Shardlow, Lexical complexity prediction: An overview, ACM Comput. Surv. 55 (2023).

[9] L. Bereska, Z. Tzifa-Kratira, R. Samavi, S. Gavves, Superposition as lossy compression — measure with sparse autoencoders and connect to adversarial vulnerability, Transactions on Machine Learning Research (2025).

[10] Y. Arase, S. Uchida, T. Kajiwara, CEFR-based sentence difficulty annotation and assessment, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 6206–6219.

[11] F. Alva-Manchego, L. Martin, A. Bordes, C. Scarton, B. Sagot, L. Specia, ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4668–4679.

[12] M. Shardlow, M. Cooper, M. Zampieri, CompLex — a new corpus for lexical complexity prediction from Likert Scale data, in: Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI), Marseille, France, 2020, pp. 57–62.

[13] S. Biderman, H. Schoelkopf, Q. G. Anthony, H. Bradley, K. O'Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff, A. Skowron, L. Sutawika, O. van der Wal, Pythia: A suite for analyzing large language models across training and scaling, in: International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA, volume 202 of *Proceedings of Machine Learning Research*, PMLR, 2023, pp. 2397–2430.

[14] Y. Belinkov, Probing classifiers: Promises, shortcomings, and advances, Comput. Linguistics 48 (2022) 207–219.

[15] C. Davis, C. Bryant, A. Caines, M. Rei, P. Buttery, Probing for targeted syntactic knowledge through grammatical error detection, in: Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 360–373.

[16] E. Gjinika, N. Arici, L. Putelli, A. E. Gerevini, I. Serina, An analysis on how pre-trained language models learn different aspects, in: Proceedings of the 5th Italian Workshop on Explainable Artificial

Intelligence, co-located with AIxIA 2024, Bolzano, Italy, November 26-27, 2024, volume 3839 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 28–41.

[17] E. Liu, G. Neubig, Are representations built from the ground up? an empirical examination of local composition in language models, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 9053–9073.

[18] N. Arici, L. Putelli, E. Gjinika, I. Serina, A. Gerevini, Learning trajectories of figurative language for pre-trained language models, in: Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025, Association for Computational Linguistics, 2025, pp. 14440–14461.

[19] L. Gao, T. D. la Tour, H. Tillman, G. Goh, R. Troll, A. Radford, I. Sutskever, J. Leike, J. Wu, Scaling and evaluating sparse autoencoders, 2024.

[20] Z. He, W. Shu, X. Ge, L. Chen, J. Wang, Y. Zhou, F. Liu, Q. Guo, X. Huang, Z. Wu, Y.-G. Jiang, X. Qiu, Llama scope: Extracting millions of features from llama-3.1-8b with sparse autoencoders, 2024.

[21] T. Lieberum, S. Rajamanoharan, A. Conmy, L. Smith, N. Sonnerat, V. Varma, J. Kramar, A. Dragan, R. Shah, N. Nanda, Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2, in: Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, Association for Computational Linguistics, Miami, Florida, US, 2024, pp. 278–300.

[22] L. Seiffe, F. Kallel, S. Möller, B. Naderi, R. Roller, Subjective text complexity assessment for German, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 707–714.

[23] S. Gooding, E. Kochmar, S. M. Yimam, C. Biemann, Word complexity is in the eye of the beholder, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 4439–4449.

[24] A. Branco, Computational complexity of natural languages: A reasoned overview, in: Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing, Association for Computational Linguistics, Santa Fe, New-Mexico, 2018, pp. 10–19.

[25] D. J. Velasco, M. T. Roque, Rethinking the role of text complexity in language model pretraining, in: Proceedings of the First BabyLM Workshop, Association for Computational Linguistics, Suzhou, China, 2025, pp. 1–28.

[26] J. M. Imperial, A. Barayan, R. Stodden, R. Wilkens, R. Muñoz Sánchez, L. Gao, M. Torgbi, D. Knight, G. Forey, R. R. Jablonkai, E. Kochmar, R. J. Reynolds, E. Ribeiro, H. Saggion, E. Volodina, S. Vajjala, T. François, F. Alva-Manchego, H. Tayyar Madabushi, UniversalCEFR: Enabling open multilingual research on language proficiency assessment, in: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Suzhou, China, 2025, pp. 9703–9755.

[27] A. Rogers, O. Kovaleva, A. Rumshisky, A primer in BERTology: What we know about how BERT works, Transactions of the Association for Computational Linguistics 8 (2020) 842–866.

[28] A. Miaschi, D. Brunato, F. Dell'Orletta, G. Venturi, Linguistic profiling of a neural language model, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 745–756.

[29] M. Fayyaz, E. Aghazadeh, A. Modarressi, H. Mohebbi, M. T. Pilehvar, Not all models localize linguistic knowledge in the same place: A layer-wise probing on BERToids' representations, in: Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 375–388.

[30] O. Skean, M. R. Arefin, D. Zhao, N. N. Patel, J. Naghiyev, Y. LeCun, R. Shwartz-Ziv, Layer by layer: Uncovering hidden representations in language models, in: Forty-second International Conference on Machine Learning, 2025.