Vector Norms as an Approximation of Syntactic Complexity

Adam Ek **Nikolai Ilinykh**
Centre for Linguistic Theory and Studies in Probability (CLASP)
Department of Philosophy, Linguistics and Theory of Science (FLoV)
University of Gothenburg, Sweden
{adam.ek,nikolai.ilinykh}@gu.se

Abstract

Internal representations in transformer models can encode useful linguistic knowledge about syntax. Such knowledge could help optimise the data annotation process. However, identifying and extracting such representations from big language models is challenging. In this paper we evaluate two multilingual transformers for the presence of knowledge about the syntactic complexity of sentences and examine different vector norms. We provide a fine-grained evaluation of different norms in different layers and for different languages. Our results suggest that no single part in the models would be the primary source for the knowledge of syntactic complexity. But some norms show a higher degree of sensitivity to syntactic complexity, depending on the language and model used.

1 Introduction

One of the most successful recent developments in NLP is the self-attention mechanism (Cheng et al., 2016; Lin et al., 2017), which has been used as the underlying operation of recent transformer models (Vaswani et al., 2017). The success of the transformer models has been wide-spread, from semantic (Tenney et al., 2019b) and syntactic (Raganato and Tiedemann, 2018; Vig and Belinkov, 2019; Clark et al., 2019) tasks, to more pragmatically focused tasks (Ettinger, 2020) and multimodal problems (Bugliarello et al., 2021). In this paper we contribute to the research on what makes transformers so successful in learning linguistic knowledge and examine the ability of such models to estimate *syntactic complexity of sentences*.

Knowing if a model reacts to the syntactic complexity of sentences is useful because if complexity can be estimated without training a model (during inference time), this can be taken into account when fine-tuning the model, allowing for more efficient sampling of batches during training (Zhao and Zhang, 2015; Katharopoulos and Fleuret, 2018), curriculum learning strategies (Bengio et al., 2009; Hacohen and Weinshall, 2019) and possibly as an estimator of unsupervised/zero-shot performance. Another aspect we consider is whether the syntactic complexity of sentences can be captured in different languages or if it works particularly well for some, as this would allow vector norms to be used as an analysis tool regardless of the language.

We examine the transformer representations for presence of linguistic knowledge by first extracting **vector norms**, as these can be obtained by simply passing a sentence through a pre-trained model without fine-tuning. The norms of vectors have been used to both analyse models (Kobayashi et al., 2020) and improve models, for example by using the norm as an indicator of the difficulty of translating a sentence (Platanios et al., 2019; Liu et al., 2020), or as a way to select informative examples from a corpora (Lu and Zhang, 2022). The L2-norm was also exploited by (Hewitt and Manning, 2019) to construct dependency trees from the dot-product of word pairs. In this paper we investigate whether the Euclidean norm (L2-norm) is an indicator of the syntactic complexity in dependency trees. We consider two distance metrics: dependency and hierarchical distance, and we investigate three sources of L2-norms: CLS, ROOT and average over tokens. We finally evaluate the ability of the transformer to estimate syntactic complexity of a sentence by looking at the pearson ρ correlation between vector norms and distance metrics of dependency trees. This will show whether the norm of the representations in a transformer increase or decrease as the syntactic complexity changes.

We present an analysis across *many languages*, and we use only pre-trained large language model representations. This is useful for many reasons, one of them being that it can help inform annotators that a certain sentence is a syntactically complex one, or not, and this can helps annotators recognize difficult sentences during the annotation process. Additionally, this work provides another use case for LLMs, namely as an aid for dataset creators. That is, if a silver-standard is obtained from a model, we still want to identify possible problematic sentences that should be reviewed by a human annotator. By analysing different norms and syntactic complexity in different languages, we also provide valuable information about how the MLM objective of transformer models encode syntactic complexity.

In this work we focus on the following questions:

- Which norms are most efficient for estimating dependency and hierarchical distance?
- Are different models better at estimating dependency or hierarchical distance?
- Is the estimation of dependency and hierarchical distance influenced by the language in question?

The questions we investigate are relevant for better interpretability of neural language models (Belinkov and Glass, 2019). We provide insights on whether transformers can be used (without training) as the knowledge source for more efficient subsequent annotation, training and fine-tuning. We also examine the differences between norms in different model layers concerning many languages.

2 Dependency distance metrics

In this paper we consider syntactic complexity of sentences through the lens of dependency trees. Quantifying these dependencies heavily relies on the word order, which can be represented either *linearly* or *hierarchically* (structurally). Depending on the language, either of the ways is preferred by its speakers and a reciprocal relation between the two can be observed, e.g. Czech relies on structural order more than English for longer sentences (Jing and Liu, 2015a). We believe that such fine-grained difference between how languages use syntax is important for the models to capture. Therefore, we represent the syntactic complexity of sentences by calculating mean linear dependency distance (MDD) and mean hierarchical dependency distance (MHD). For evaluation we compute correlation between different vector norms and MDD / MHD and treat this intrinsic measure as an indicator of the amount of linguistic knowledge about syntactic complexity that the model encodes. We extract dependency trees from the Universal Dependency treebanks (de Marneffe et al., 2021). For a tree such as Figure 1, there are two distance metrics that we are interested in, *dependency* and *hierarchical* distance.

Mean Dependency Distance (MDD) In a dependency tree, the mean dependency distance (Liu, 2008) is the number of intervening words between the head, and the dependent. We can consider a function f_{head} that takes as input a word, and outputs the distance to its head. To calculate the mean dependency distance, we employ Equation (1):

$$\mathsf{MDD}(S) = \frac{1}{n}\sum_{i=1}^{n} f_{head}(s_i) \tag{1}$$

For example, in Figure 1, the only head-dependent pair with more than one intervening word is that between ROOT and *runs*, where there are 2 intervening words. Thus, the mean dependency distance of the sentence is $\frac{2+1+1+1+1}{5} = 1.2$.

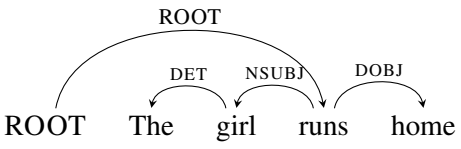


Figure 1: Dependency tree of the sentence *The girl runs*.

Mean Hierachical Distance (MHD) For calculating mean hierarchical distance (Jing and Liu, 2015b), we consider the shortest path from word i to the ROOT node as a function f_{root}. We calculate the mean hierarchical distance as follows:

$$\mathsf{MHD}(S) = \frac{1}{n}\sum_{i=1}^{n} f_{root}(s_i) \tag{2}$$

The hierarchical distance between words in the sentence "The girl runs home" is visualised in Fig-

ure 2, where we arrive at an average hierarchical distance of $\frac{1+2+2+3}{4} = 2$. Because the ROOT is always distance 0 from itself, we ignore this in the calculations.

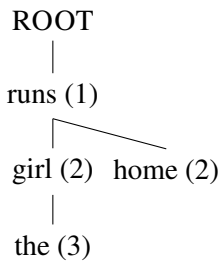


Figure 2: Hierarchical representation of Figure 1. The hierarchical depth for each word is given in parentheses.

3 Vector norms

We are interested in whether the vector norms capture behavior trends of syntactic complexity of sentences in different languages. We use the euclidean (L_2) norm as our primary method for norm computation. The euclidean norm is defined as follows over a vector v of length n:

$$\|v\|_2 = \sqrt{v_0^2 + \ldots + v_n^2} \tag{3}$$

The representations that are used to compute the sentence vector norm can be taken arbitrarily from different representations inside the model. Here we extract *three* such representations and compute norms for each of them:

- **CLS**: the norm of the CLS token which is appended to every sentence, and which functions as a sentence "summary" in the transformer model. The CLS norm of the first self-attention layer in the model is the same for all sentences, thus do not provide any insights.
- **ROOT**: the norm of the ROOT token. In dependency trees this token represents the top-level node in the tree. Because transformers use sub-word tokenisation the root may contain several sub-word tokens. In this case we consider the mean representation of all sub-words.
- **MEAN**: the mean of the norms of all tokens in a sentence. This norm would then consider the average representation across all words in the sentence.

In our experiments we consider the norms obtained from two different models, multilingual BERT (m-BERT) (Devlin et al., 2019) and XLM-RoBERTa (XLM-R) (Conneau et al., 2020). For both models, we use the base model provided by the HuggingFace (Wolf et al., 2019) library.

We compute the Pearson correlation (Freedman et al., 2007) between the dependency distance metrics and the vector norms. We obtain a *mean correlation score* for each language by averaging scores across all sentences in this language. A high correlation score means that a specific type of vector norm within the particular model's layer reflects bigger distance scores. A low score (or a negative score) indicates that the vector norm and dependency distance are in an opposite trend: when one becomes high, another one gets low. Observing differences in the behaviour of different vector norm representations and distance metrics allows us to determine which mean will likely encode which type of knowledge.

4 Treebanks

We use the parallel sentences from the PUD corpus (Zeman et al., 2017). For each language in the dataset, it contains 1000 sentences. The sentences are the same for all languages and have been translated and annotated by experts. In the PUD dataset, dependency trees from the following languages are included: Arabic, Czech, English, French, German, Hindi, Icelandic, Italian, Polish, Portuguese, Russian, Spanish, Swedish, Thai, Chinese, Turkish, Korean, Japanese, Indonesian and Finnish. The majority of the languages are from the Indo-European language family, however, other distant families are also included such as Uralic, Turkic, Austronesian, Sino-Tibetan, and Tai-Kadai. The primary attractive feature of the PUD dataset for our experiments is that the results we obtain for the different languages are directly comparable because the same sentence is translated. This allows us to reliably ascertain the models ability to encode different dependency tree metrics across languages.

We summarize each language's mean dependency and hierarchical distances in Table 1. The hierarchical distance for all languages is higher than the linear distance between syntactic dependants. This artefact of the dataset is crucial as it might affect how successful different norms are in capturing different dependencies. Previous work

Language	MDD	MHD
Arabic	3.01	4.47
Chinese	3.29	4.20
Czech	3.01	4.29
English	3.16	4.23
Finnish	2.83	3.99
French	3.09	4.43
German	3.71	4.21
Hindi	3.51	4.33
Icelandic	2.86	4.22
Indonesian	2.84	4.27
Italian	3.09	4.41
Japanese	2.87	4.51
Korean	2.54	4.40
Polish	2.87	4.26
Portuguese	3.07	4.41
Russian	2.91	4.31
Spanish	3.07	4.42
Swedish	3.03	4.15
Thai	2.38	4.61
Turkish	2.70	4.19
Mean	2.99	4.31
Std	0.74	0.70

Table 1: Mean dependency and hierarchical distance for different treebanks.

has shown that models can capture hierarchical structures in natural language, but only to a degree (Wilcox et al., 2019); therefore, it would be interesting to see whether the differences in distances found in the dataset have an impact on the correlation scores. The standard deviation reveals that both metrics vary about the same amount. The distribution of mean dependency and hierarchical distances over all treebanks is given in Figure 3. The median of dependency distance is 2.92 with a skew of 0.74, while the mean hierarchical distance is 4.25 with a skew of 0.73. Because the skew is below 1 for both distances, this shows that, when considering the distances over all languages, they tend to minimize both the MDD, and MHD (Futrell et al., 2015).

5 Results

Table 2 and Table 3 show Pearsons ρ correlation scores between various norms and the distance metrics for m-BERT and XLM-R respectively. On a high level, we observe that m-BERT and XLM-R exhibit vastly different behaviours regarding what norms best predict the distances. For

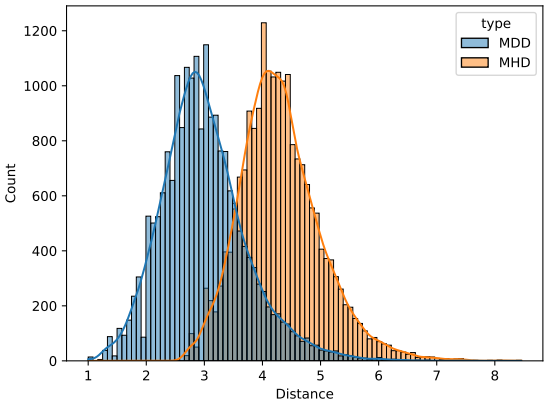


Figure 3: Distribution of MDD and MHD over all PUD treebanks.

base m-BERT (Table 2), we see that the mean norm over all sub-word tokens in a sentence correlates strongly with both distance metrics. That is, if this specific norm of a sentence in m-BERT is high, it is likely that either the MDD or MHD is also high. Based on the mean correlation score, we see that MEAN is generally a better predictor than the other two norms (ROOT and CLS). A similar result is observed for the base XLM-R model (Table 3): the MEAN norm has a positive correlation with either of the distance metrics. However, this trend is not strongly pronounced since the mean correlation scores for the best norm are lower for XLM-R (0.18 and 0.19) than for m-BERT (0.54 and 0.62). The CLS norm exhibits a stronger negative correlation for XLM-R models (in fact, it's the best-observed correlation for XLM-R), which means that when the norm is high, the mean distance between the words is small. ROOT-based vectors are the least useful across both models. The results indicate that MEAN norms in m-BERT are better estimators for sentences' syntactic complexity than XLM-R, which has a higher correlation for CLS norms, although in a different direction. However, XLM-R significantly outperforms m-BERT on various cross-lingual tasks (Conneau et al., 2020); still, its internal representations are more diluted, and there is no clear correspondence to the trends in syntactic complexity of sentences because of overall slightly weaker correlations.

	Root	Mean	CLS
Arabic	0.17	0.53	-0.20
Chinese	0.43	0.49	0.04*
Czech	0.16	0.53	-0.19
English	0.22	0.56	-0.10
Finnish	0.22	0.49	-0.16
French	0.27	0.62	-0.10
German	0.26	0.58	-0.11
Hindi	0.38	0.60	-0.16
Icelandic	0.20	0.49	-0.19
Indonesian	0.28	0.56	-0.20
Italian	0.35	0.56	-0.06*
Japanese	0.36	0.54	0.08
Korean	0.09	0.50	-0.06*
Polish	0.12	0.54	-0.15
Portuguese	0.30	0.60	-0.19
Russian	0.23	0.57	-0.14
Spanish	0.33	0.57	-0.13
Swedish	0.29	0.55	-0.23
Thai	0.06	0.43	-0.15
Turkish	-0.00*	0.54	0.02
Mean	0.23	0.54	-0.11

(a) Dependency distance

	Root	Mean	CLS
Arabic	0.21	0.57	-0.28
Chinese	0.53	0.57	0.04*
Czech	0.21	0.66	-0.15
English	0.33	0.64	-0.06
Finnish	0.30	0.66	-0.20
French	0.35	0.65	-0.14
German	0.32	0.66	-0.22
Hindi	0.40	0.61	-0.21
Icelandic	0.23	0.61	-0.12
Indonesian	0.33	0.62	-0.20
Italian	0.38	0.66	-0.16
Japanese	0.45	0.61	-0.09
Korean	0.08	0.54	-0.14
Polish	0.18	0.61	-0.14
Portuguese	0.38	0.67	-0.19
Russian	0.25	0.65	-0.13
Spanish	0.34	0.65	-0.13
Swedish	0.36	0.66	-0.19
Thai	0.18	0.58	-0.25
Turkish	0.02*	0.62	-0.07
Mean	0.29	0.62	-0.15

(b) Hierarchical distance

Table 2: **base m-BERT**: Pearsons ρ between the CLS, ROOT, and MEAN with respect to MDD (a) and MHD (b) distances extracted from the PUD treebanks. Correlations with $p > .05$ are indicated by *.

6 Peeking inside the model's layers

We next perform a more fine-grained analysis of the internal models' representations. Our goal is to examine to what degree specific layers of the models capture information about the syntactic complexity of sentences. This analysis allows us to narrow down the search for the source of better representations in large language models. The correlations per layer using the m-BERT model are shown in Figure 4 for MDD and in Figure 5 for MHD. Similarly, Figure 6 and Figure 7 show correlations per layer for the XLM-R model.

Discussion For the m-BERT model, MEAN vector norms have large correlation scores in the first few layers, while in deeper layers, the correlation drops only to spike again in the last layer. This behaviour is approximately identical for all languages, with some having a slightly smaller correlation with MDD across all layers, e.g. Thai. The trend is very similar for correlation with MHD: the scores are the highest in approximately the first six layers and smaller in deeper layers except for the last layer, in which the scores are high again. This result mirrors previous findings showing that language models capture syntactic knowledge in earlier layers, which allows them to learn complementary semantic knowledge in deeper layers (Peters et al., 2018; Kovaleva et al., 2019; Tenney et al., 2019a), which could be the reason for a more negligible correlation with syntactic complexity in deeper layers. CLS vector norms in the first layer have the highest (positive) correlation with both types of distances across most languages. The correlation is still strong in subsequent layers, although in a different direction (negative), with a substantial reduction in the last layer. Some notable exceptions are Chinese, Japanese and Thai, for which the highest correlation is observed in the second or the third layer. After these layers, the correlation effect is smaller than for other languages. ROOT vector norms have a small correlation with the syntactic complexity, suggesting that these norms are not informative enough and richer representations are required.

For the XLM-R, correlation scores for MEAN vector norms vary a lot among the layers. However, the correlation spikes in the second layer

	Root	Mean	CLS
Arabic	0.22	0.06*	-0.35
Chinese	0.29	0.24	-0.37
Czech	0.15	0.37	-0.24
English	0.19	0.15	-0.29
Finnish	0.07	0.08	-0.27
French	0.12	0.29	-0.32
German	0.16	0.08	-0.39
Hindi	0.25	0.53	0.08
Icelandic	0.10	0.21	-0.31
Indonesian	0.13	-0.10	-0.32
Italian	0.14	0.21	-0.28
Japanese	0.16	0.32	-0.33
Korean	0.17	0.09	-0.36
Polish	0.15	0.13	-0.26
Portuguese	0.12	0.03*	-0.39
Russian	0.20	0.24	-0.30
Spanish	0.15	0.10	-0.36
Swedish	0.10	0.10	-0.35
Thai	-0.06*	0.39	-0.15
Turkish	0.11	0.13	-0.32
Mean	0.14	0.18	-0.29

(a) Dependency distance

	Root	Mean	CLS
Arabic	0.16	0.04*	-0.42
Chinese	0.21	0.29	-0.46
Czech	0.19	0.44	-0.33
English	0.16	0.06	-0.44
Finnish	0.12	0.07	-0.41
French	0.12	0.28	-0.38
German	0.18	0.09	-0.43
Hindi	0.22	0.46	-0.11
Icelandic	0.11	0.23	-0.38
Indonesian	0.11	-0.14	-0.41
Italian	0.14	0.19	-0.41
Japanese	0.06	0.41	-0.42
Korean	0.26	0.12	-0.45
Polish	0.11	0.10	-0.40
Portuguese	0.08	0.00*	-0.46
Russian	0.24	0.25	-0.42
Spanish	0.17	0.08	-0.46
Swedish	0.13	0.09	-0.45
Thai	0.09	0.53	-0.27
Turkish	0.09	0.12	-0.37
Mean	0.14	0.19	-0.39

(b) Hierarchical distance

Table 3: **base XLM-R**: Pearsons ρ between the CLS, ROOT, and MEAN with respect to MDD (a) and MHD (b) distances extracted from the PUD treebanks. Correlations with $p > .05$ are indicated by *.

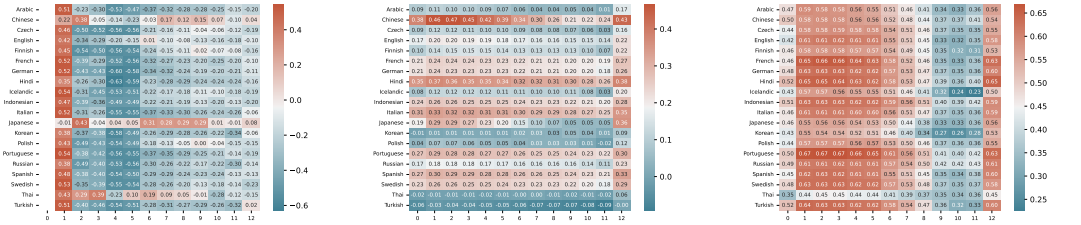
and reaches low values in the last layer for many languages, which is opposite to the behaviour of mean vector norms in m-BERT. CLS vector norms and their correlation with distance metrics similarly varies between layers, but not languages. All languages appear to have similar correlation scores depending on the layer. ROOT vector norms have a very small correlation (positive or negative) for all languages across all layers except the last one. Interestingly, XLM-R does not encode the knowledge of syntactic complexity in representations of the root of the sentence, which mirrors results for m-BERT.

Overall, m-BERT exhibits much more pronounced differences between layers regarding the syntactic complexity of sentences. XLM-R, on the contrary, is much harder to understand and use for finding parts in the model where a particular norm is the most useful. For m-BERT, we recommend using mean vector norms from the first layers to identify the knowledge of syntactic complexity. For XLM-R, CLS or MEAN can be used. Neither model has useful ROOT representations.

7 Discussion

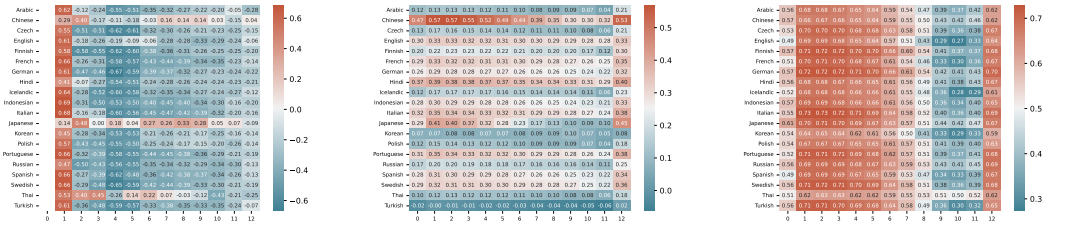
Norm types and models We find that the ROOT and MEAN norms produce only weak correlations in the XLM-R model across languages, suggesting that XLM-R better encodes syntactic information about sentences in its CLS token rather than in the sub-word tokens themselves. We see an inverse trend for m-BERT: syntactic complexity is strongly encoded in the sub-word tokens (MEAN and ROOT) while not in the CLS token. This indicates that concerning encoding syntactic complexity, the XLM-R model performs well in pooling SC information to the CLS token during the language modelling training and also reduces indications of syntactic complexity from the sub-word tokens, whereas in m-BERT syntactic complexity is strongly encoded in the sub-word tokens. We suggest that the CLS token in XLM-R generally contains more information about syntax than the CLS token in m-BERT.

MDD vs MHD In both models all three vector norms have a stronger correlation with MHD and a weaker correlation with MDD. However, the mod-



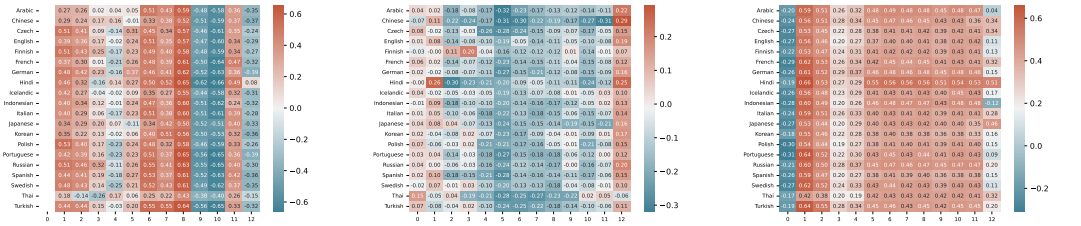
(a) CLS (b) Root (c) Mean

Figure 4: **base m-BERT**: correlation with MDD across layers.



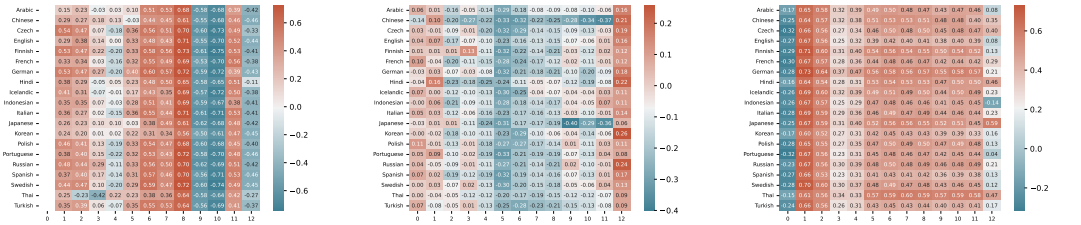
(a) CLS (b) Root (c) Mean

Figure 5: **base m-BERT**: correlation with MHD across layers.



(a) CLS (b) Root (c) Mean

Figure 6: **base XLM-R**: correlation with MDD across layers.



(a) CLS (b) Root (c) Mean

Figure 7: **base XLM-R**: correlation with MHD across layers.

els have not been provided with information about the hierarchy of dependencies in sentences; therefore, it is surprising to see such a trend. We hypothesise that this is because of how self-attention functions. In self-attention, each word in a sentence is multiplied by all other words in this sentence. This process leads to the model having no inherent inductive bias towards linear relationships; rather, it enables models to find relationships which are not linear more easily. Therefore, this directly impacts the differences between how linear and hierarchical knowledge is acquired.

Previous work has shown that self-attention might capture hierarchically organised knowledge in different kinds of tasks (Yang et al., 2019; Ilinykh and Dobnik, 2021); the result aligns with our finding.

Language-to-language comparison Knowing which norms better correlate with which languages across models is helpful because this provides insights into what type of representations one should extract when working with specific languages. Based on the results in Section 5, for m-BERT, the MEAN norm correlates with syntactic complexity the most across languages. We can also note that m-BERT is better correlated with the MHD across languages and norms. This is also the case for the ROOT and MEAN norm, while the CLS norm exhibits some deviations for this. For the CLS norm: Czech, English, Icelandic, Japanese, Russian and Swedish show a slightly stronger negative correlation with MDD than MHD. There appears to be no clear explanation, as the languages differ in the script used (Latin, Cyrillic, and Kanji) and the language families. We do not put much weight on this deviation because the CLS norm generally provides a weak negative correlation with both distances.

XLM-R shows more varied trends. The CLS norm shows the strong (negative) correlation with MDD and MHD, $-.29$ and $-.39$, respectively. As in m-BERT, this correlation is stronger for MHD. As both models are based on the self-attention mechanism, this is not a surprising finding. However, XLM-R exhibits some stronger deviations for the preference towards MHD. This happens mainly for the root token, where Arabic, Chinese, English, Hindi, Japanese, Polish, Portuguese and Turkish better encode MDD. Interestingly, we see this preference in so many languages for the root token, as it is the top node in the dependency tree. This indicates that while the ROOT norm is not the most correlated norm on average, it exhibits more variation in what it captures than the other norms. MDD is preferred only for Hindi with the CLS token for the other norms and for French, Hindi, Italian and Polish using the MEAN norm.

Overall, there are differences in how m-BERT and XLM-R capture the knowledge of the syntactic complexity of sentences across languages. While m-BERT is computing better representations for mean norm over sub-words for many of the languages in the PUD treebank, XLM-R has better CLS-token representations, possibly due to the larger dataset that it has been pre-trained on.

7.1 Research questions revisited

We can now consider the initial research questions posed in the introduction.

Which norms are most efficient for estimating dependency and hierarchical distance? In general, for m-BERT the mean norm over the tokens in the sentence is the best indicator across languages. For XLM-R the story is different, with the mean norm over tokens not showing a considerable correlation. In contrast, the CLS-based norms show a stronger negative correlation with both dependency and hierarchical distance.

Are different models better at estimating dependency or hierarchical distance? Yes, we find that both m-BERT and XLM-R exhibit stronger correlations with hierarchical (structural) distances than dependency (linear) distances, on average. However, we can also observe deviations from this, both concerning the models and norm type.

Is the estimation of dependency and hierarchical distance influenced by the language in question? We can note that we can observe quite different correlations for all vector norms when comparing the languages; however, they still generally follow the same trends for a given norm type.

8 Conclusions and future work

We conclude by hypothesising how our findings could aid data annotation efforts. We show that for both models, we can identify a specific type of norm that is a relatively strong indicator of syntactic complexity across languages. We see at least two potential uses of such a result. First, it enables dataset creators to use a machine learning model to label data and then select examples which may require the aid of a human to fine-tune the annotation. Secondly, our findings enable dataset creators to rank examples in terms of syntactic complexity. This can be used to assign some examples to annotators based on their experience, where expert annotators are assigned one set of examples, and novice/intermediate annotators are assigned another set of examples. For future work, we would like to explore using vector norms for these purposes, as it can help create better datasets by not assigning examples to annotate randomly but in an informed manner.

Acknowledgements

We would like to thank the reviewers for their helpful comments. The research reported in this paper was supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

References

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

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, volume 382 of *ACM International Conference Proceeding Series*, pages 41–48. ACM.

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2021. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language BERTs. *Transactions of the Association for Computational Linguistics*, 9:978–994.

Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long short-term memory-networks for machine reading. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 551–561, Austin, Texas. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

David Freedman, Robert Pisani, and Roger Purves. 2007. Statistics (international student edition). *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proc Natl Acad Sci U S A*, 112(33):10336–10341.

Guy Hacohen and Daphna Weinshall. 2019. On the power of curriculum learning in training deep networks. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2535–2544. PMLR.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Nikolai Ilinykh and Simon Dobnik. 2021. What does a language-and-vision transformer see: The impact of semantic information on visual representations. *Frontiers in Artificial Intelligence*, 4.

Yingqi Jing and Haitao Liu. 2015a. Mean hierarchical distance augmenting mean dependency distance. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 161–170, Uppsala, Sweden. Uppsala University, Uppsala, Sweden.

Yingqi Jing and Haitao Liu. 2015b. Mean hierarchical distance augmenting mean dependency distance. In *Proceedings of the third international conference on dependency linguistics (Depling 2015)*, pages 161–170.

Angelos Katharopoulos and François Fleuret. 2018. Not all samples are created equal: Deep learning with importance sampling. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2530–2539. PMLR.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7057–7075. Association for Computational Linguistics.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of bert. *arXiv preprint arXiv:1908.08593*.

Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159–191.

Xuebo Liu, Houtim Lai, Derek F. Wong, and Lidia S. Chao. 2020. Norm-based curriculum learning for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 427–436. Association for Computational Linguistics.

Yu Lu and Jiajun Zhang. 2022. Norm-based noisy corpora filtering and refurbishing in neural machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5414–5425.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Matthew E. Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium. Association for Computational Linguistics.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabás Póczos, and Tom M. Mitchell. 2019. Competence-based curriculum learning for neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1162–1172. Association for Computational Linguistics.

Alessandro Raganato and Jörg Tiedemann. 2018. An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 287–297, Brussels, Belgium. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4593–4601. 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, and Ellie Pavlick. 2019b. What do you learn from context? probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy. Association for Computational Linguistics.

Ethan Wilcox, Roger Levy, and Richard Futrell. 2019. Hierarchical representation in neural language models: Suppression and recovery of expectations. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 181–190, Florence, Italy. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Baosong Yang, Longyue Wang, Derek F. Wong, Lidia S. Chao, and Zhaopeng Tu. 2019. Assessing the ability of self-attention networks to learn word order. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3635–3644, Florence, Italy. Association for Computational Linguistics.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis M. Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajic Jr., Jaroslava Hlavácová, Václava Kettnerová, Zdenka

Uresová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çagri Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali El-Kahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. Conll 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Vancouver, Canada, August 3-4, 2017*, pages 1–19. Association for Computational Linguistics.

Peilin Zhao and Tong Zhang. 2015. Stochastic optimization with importance sampling for regularized loss minimization. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1–9. JMLR.org.