Entrenchment Matters: Investigating Positional and Constructional Sensitivity in Small and Large Language Models

Bastian Bunzeck and **Sina Zarrieß**
Computational Linguistics
Bielefeld University, Germany
{bastian.bunzeck, sina.zarriess}@uni-bielefeld.de

Abstract

The success of large language models (LMs) has also prompted a push towards smaller models, but the differences in functionality and encodings between these two types of models are not yet well understood. In this paper, we employ a perturbed masking approach to investigate differences in token influence patterns on the sequence embeddings of larger and smaller RoBERTa models. Specifically, we explore how token properties like position, length or part of speech influence their sequence embeddings. We find that there is a general tendency for sequence-final tokens to exert a higher influence. Among part-of-speech tags, nouns, numerals and punctuation marks are the most influential, with smaller deviations for individual models. These findings also align with usage-based linguistic evidence on the effect of entrenchment. Finally, we show that the relationship between data size and model size influences the variability and brittleness of these effects, hinting towards a need for holistically balanced models.

1 Introduction

Recent years have witnessed an exponential growth in the size of language models, which has led to significant improvements in their performance on various natural language processing tasks. However, the reasons behind the remarkable success of LMs remain elusive, and it is questionable whether further growth will enhance their performance (Hong et al., 2022). More recently, it has been shown that small models can potentially learn linguistic structure equally well (Warstadt et al., 2020b; Zhang et al., 2021; Huebner et al., 2021). Because of neural network's opaque functionality, the reasons for these similarities and differences are not yet well understood. While grammatical evaluation suites (Warstadt et al., 2020a; Huebner et al., 2021; Newman et al., 2021) focus more on model output, evaluation approaches from the field of BERTology (Rogers et al., 2020) try to address this problem by studying the model's internal representations and mechanics. The present paper employs a perturbed masking approach (Wu et al., 2020) to study the influence of syntactic and constructional factors on lexical influence in sentence embeddings, and their differences between smaller and larger models.

To investigate these differences, we propose an approach inspired by usage-based linguistics. In the usage-based view, grammar is seen as emerging from language use and domain-general learning mechanisms (Diessel, 2019). Constructions, form-meaning pairings on all levels of linguistic analysis, are seen as the essential building blocks of language (Fillmore, 1988; Goldberg, 2003). Domain-general processes that influence such construction grammars are highly dependent on frequency effects in the input. For example, repeated use of a linguistic structure leads to it becoming more entrenched, more *unit-like*, in a speaker's "cognitive organization" (Langacker, 1987, 59). As artificial neural networks are domain-agnostic, statistical learners that create their linguistic systems through repeated use in the learning process, such effects attested in human language users should also be present in artificial learners. Consequently, usage-based approaches should be able to provide new insights on understanding the linguistic capabilities of language models, and the differences between large and small models (with different amounts of input) in particular.

Within our usage-based framework, we explore the influence that individual tokens have on the embeddings of their sequence. In opposition to grammatical test suites that challenge LLMs' abilities on very specific phenomena and structures, we aim to explore and analyze the linguistic abilities of LMs in terms of general positional and constructional factors and influences in their representations. Furthermore, by comparing these aspects for models trained on different amounts of

Proceedings of the 2023 CLASP Conference on Learning with Small Data, pages 25–37
September 11–12, 2023.

linguistic data, we aim to find out whether, and if so, how fast constructional entrenchment and generalizations may arise. Because their representations are shaped by less input data, smaller models may exhibit less entrenchment effects, and be thus more brittle and sensitive. By comparing the most influential parts of speech for the models and construction types, we aim to find out on which grammatical categories the sequence embeddings depend the most and whether this is changed through the amount of training data. Finally, the frequency and information effects should affect the representations diametrically. The present analysis will tell if any effect is stronger in larger LMs.

Generally, we find that LMs of all sizes have a bias towards attributing more weight/influence to tokens at the end of a sentence, which aligns with the information-driven aspect of the linguistic theory. However, although there are construction-dependent differences, this effect does not vary systematically between the construction types. This is surprising, considering their differences in lexical specificity. Furthermore, we find that this bias is influenced by model size, but not in a linear fashion. Finally, we find that all models assign the highest importance to lexical words, especially nouns. In this sense, our results suggest that entrenchment as a property of statistical learners does indeed map from usage-based theories to artificial learners. Yet, its interplay with model structure and learning processes remains complex and not completely transparent.

2 Motivation for usage-based approaches to LMs

Established evaluation suites for grammatical abilities (Warstadt et al., 2020a; Huebner et al., 2021; Newman et al., 2021) often work by focusing on models' preferences for grammatical utterances over their ungrammatical counterparts. These techniques are inspired by a rather strict generative view of language, which assumes a pre-endowed human language faculty that generates grammatical strings of words from a hypothesized mental hierarchical structure (Chomsky, 1957, 1965). If the goal of neural language modelling was to recreate this, then only testing on phenomena like binding or filler-gap relations would be sufficient. However, alternative approaches to linguistic theory question these notions. The usage-based approach sees grammar as a fuzzy mental model of language that is constantly shaped and re-shaped by domain-general cognitive mechanisms, such as automatization, entrenchment or analogy, through input and usage (Tomasello, 2003; Diessel, 2019). The resulting mental representations in the form of linguistic constructions are influenced by frequency effects. For example, forms that are perceived and produced more often are more deeply entrenched in mental grammar (Schmid, 2015). On the syntactic level, such effects have syntagmatic and paradigmatic dimensions. The syntagmatic dimension refers to which elements occur sequentially, whereas the paradigmatic dimensions is concerned with the variation possible for certain positions/slots in syntagms. Constructions exhibit different levels of such variation. For example, wh-questions like *Where is the butter?* only have a limited number of options for the question word (first slot) or the auxiliary (second slot), whereas the last slot can be filled by any noun. These variation effects manifest in different phenomena. Research from child-directed speech shows that spoken language is organized around *lexical frames*, lexically restricted sentence beginnings that occur with a much higher frequency than their lexically diverse counterparts, and which differ for syntactic construction types (Cameron-Faulkner et al., 2003). While the possible linguistic variation in language production is quasi-infinite, speakers rely on highly frequent, mentally automatized combinations to initiate utterances. This preference is commonly related to ease of production, a factor that is also realized in phonetic reduction or syntactic contraction of high frequency units (Bybee and Thompson, 1997). Such producer-oriented factors grounded in automatization and entrenchment are, however, not the only usage-based variables shaping variation in utterance production. For example, the information weight principle (Behaghel, 1930; Quirk, 1972; Arnold et al., 2000) posits that new information and longer, "heavier" constituents in English are commonly placed at the end of utterances, which facilitates communicative ease from a hearer-oriented perspective. Such aspects only play a very minor role in current approaches to evaluating the grammatical abilities and behaviour of LMs, although they share many underlying concepts these models. Consequently, a new, usage-based paradigm to the evaluation and analysis of LLMs is needed, as it enables new insights that are not derivable from current approaches.

3 Related work

Although perturbed masking, the analysis of model architecture or size, and contructionist/usage-based approaches to NLP have had little to no overlap in previous research, they have been used for insightful analyses on their own. The following paragraphs offer a short review of the current literature in these research directions.

Perturbed masking Wu et al. (2020) show that perturbed masking can be used to retrieve dependency trees, constituency trees and document-level discourse structures by inducing tree structures from influence matrices, based on tokens connected by higher influence values. While not as exact as other parsing approaches, they showed that BERT-based representations already encode syntactic structure. Taktasheva et al. (2021) investigate the influence of syntactic perturbations through position shifts of syntactically grouped n-grams and clauses inside sentences for English, Swedish and Russian BERT models. They find that the perturbation patterns vary for languages with different degrees of word order flexibility and that syntactic representations can be better restored from languages with fixed word order (e.g. English). In an earlier approach, the NLIZE (Liu et al., 2019a) system visualized perturbation-based changes in attention heads and output weights for natural language inference tasks. The approach has not yet been used for construction-oriented analyses or the investigation of smaller models.

Model size Differences between smaller and larger models have only begun to get systematically investigated, and existing studies have arrived at somewhat contradictory conclusions. Warstadt et al. (2020b) trained a variety of RoBERTa models with growing amounts of data, ranging from 1M to 1B tokens. They show that only larger models begin to exhibit preferences for linguistic generalizations over surface generalizations. The additional amount of data appears crucial for this difference. In contrast, the BabyBERTa model (Huebner et al., 2021) restricts the model size, number of intermediate layers and attention heads. Its training data is comparably small and was sampled from child-directed speech from the CHILDES corpora. Despite these limitations, its performance across their own evaluation suite, Zorro, is similar to the much larger RoBERTa-base model, questioning if ever larger amounts of data are actually needed, or whether the combination of hyperparameters and training data is actually responsible for emergent generalizations.

Construction grammar More recently, LLMs have also begun to be investigated from a construction grammar viewpoint. Tayyar Madabushi et al. (2020) show through a series of probing experiments that BERT embeddings already contain information that could be seen as constructionist, for example by being able to successfully determine whether two sentences with little to no lexical overlap instantiate the same grammatical construction. Tseng et al. (2022) fine-tune a BERT model for a cloze completion task on open slots in Taiwanese Mandarin constructions and show that it improves performance. Moreover, sentences that instantiate the same construction tend to be spatially closer in the vector space than sentences with different constructions but the same main verb (Li et al., 2022). However, it remains questionable how applicable such knowledge is, as Weissweiler et al. (2022) find that LLMs fail to deduce conclusions from the comparative correlative construction in an inference task. Finally, Weissweiler et al. (2023) summarize the previous line of constructionist inquiry into LLMs. They find that current research has focused on only a very limited set of constructions and that there are differences in what is assumed to be evidence for the presence of constructionist information in LLMs. Consequently, they call for a diversification of constructionist research in terms of data sources and methodology. The present paper responds to this call by investigating constructions as processing units and their influence on sequence embeddings. By employing constructions as an additional analytical factor, not as the end point of the analysis, we expand on this previous research.

4 Methods

4.1 Perturbed masking

We use Wu et al.'s (2020) perturbed masking approach to calculate the influence of a token x on its sequence. This approach is adequate as a measure of influence because it captures the global influence patterns between all token pairs in a sequence, and not only the influence of one token on the entirety of a sequence.

1. For each other token y in the sequence:
   (a) y is replaced with the <mask> token

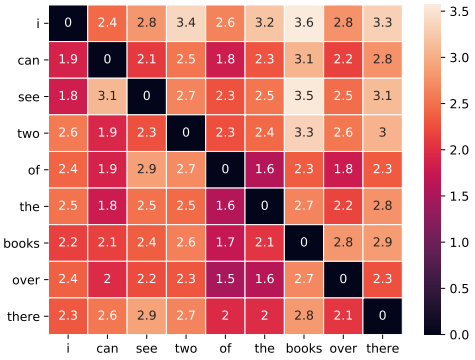

Figure 1: Influence heatmap for the sentence *I can see two of the books over there* encoded with roberta-base

(b) the sequence embedding s_y with masked y is computed

(c) x is additionally replaced with the <mask> token

(d) the embedding $s_{x,y}$ for the sequence with both x and y masked is calculated

(e) the vector distance d between s_y and $s_{x,y}$ is calculated to measure the influence of x on the embedding of y

2. By averaging the distances d between token x and all other tokens y, an average influence value of x on the sequence is acquired

We use the penultimate layer of the BERT models as the source of our embeddings, as Devlin et al. (2019) report that these embeddings perform consistently well across a variety of tasks.

By examining how certain tokens impact their embeddings more significantly, we can assess the degree to which these tokens become deeply embedded (or *entrenched*) in the learning model. Tokens with a stronger influence on their embeddings are indicative of greater entrenchment, reflecting their increased importance and resistance to modification within LMs.

4.2 Test data

We chose naturally occurring sentences from the CHILDES family of corpora (MacWhinney, 2000) as the basis of our analysis. They are especially suited to this experimental setup, because the vocabulary of child-directed speech is restricted to fairly frequent words that should be present in all models' training data, and the individual sentences are rather short and syntactically not overly complex, yet grammatical. Due to the uniform nature of child-directed speech, we also control for the influence of highly unusual or infrequent words that could disproportionately affect the perturbation data.

We sampled a data set of 3.000 test sentences from the English section of CHILDES – 1.000 sentences per construction type of interest. They were retrieved from the corpus through pattern matching on part-of-speech-tagged data. We annotated the CHILDES data with 14 different construction types inspired by Cameron-Faulkner et al. (2003) with a construction parser that operates on word class patterns. We chose three focus constructions[1] – imperatives, wh-questions and transitive sentences. We decided on these construction types because they differ in their word order (and its strictness), as well as in their lexical variation (Cameron-Faulkner et al., 2003). While transitives, for example, have a near-infinite amount of possible beginnings, wh-questions are constrained to the word class of interrogatives. This variation in syntactic and lexical patterns should shed additional light on positional and other entrenchment effects – focusing on one construction type only could taint the results by being biased from these factors.

Our parser retrieved the construction types with an accuracy of over 93% compared against a manually annotated ground-truth data set. For each construction type of interest, we then sampled 1.000 sentences randomly. To reduce variation introduced by different sentence lengths or patterns of clausal combination, only utterances with nine tokens or less were considered. The mean utterance length lies a little below that (M = 5.84, SD = 1.38), with wh-questions being the shortest (M = 5.20, SD = 1.54), transitive sentences being the longest (M = 6.28, SD = 1.25) and imperatives in between (M = 6.06, SD = 1.06).

4.3 Models

To maintain comparability of model architecture, we exclusively analysed models with RoBERTa architectures. These include the two original roberta-base[2] and roberta-large (Liu et al., 2019b) models, the distilled distilroberta-base (Sanh et al., 2020) as well as models trained with different amounts of input by Warstadt et al. (2020b). The model properties are compared in Table 1. The training data for all models was sourced from a combina-

[1]The respective patterns for the three constructions can be found in appendix A.

[2]For the rest of this paper, we denote the models by their lowercase names.

	Hidden layers	Parameters	Attention heads	Embedding size	Training data
roberta-base	12	125M	12	768	30B tokens
roberta-large	24	355M	16	1024	30B tokens
distilroberta-base	6	82M	12	768	30B tokens
roberta-med-small-1M	6	45M	8	512	1M tokens
roberta-base-10M	12	125M	12	768	10M tokens
roberta-base-100M	12	125M	12	768	100M tokens
roberta-base-1B	12	125M	12	768	1B tokens

Table 1: Hyperparameters of target models

tion of BookCorpus, English Wikipedia, CC-News, OpenWebText, and Stories dataset, and sampled down for the smaller models.

Although all of these eight models are large LMs, some of them can be considered "small" for the sake of the present analysis. roberta-med-small-1M features a reduced architecture and smaller dataset (1M tokens), whereas roberta-med-small-10M only has a smaller dataset (10M tokens) but no reduced architecture. Finally, distilroberta-base has a smaller architecture that was later fine-tuned to mirror the larger roberta-base (based on 30B tokens). Consequently, the first two models can be considered small in terms of architecture and data, whereas the latter are small in data or architecture. BabyBERTa was excluded from the analysis because its training data contains a part of the test data we used.

4.4 Experimental setup

For all 3.000 sentences and all 7 models, we perform perturbed masking with the transformers library (Wolf et al., 2020). Figure 1 shows the result of one perturbation run on the sequence *I can see two of the books over there*. The influencing words are shown on the x-axis, the influenced words on the y-axis. A higher numerical value, shown by brighter-colored cells, stands for a larger vector distance between the once and twice masked sequence embeddings. This indicates that the respective influencing word exerts a higher influence on the embedding of the influenced word – it changes its numerical values more strongly. As Wu et al. (2020) note, these patterns often align with grammatical relations. For example, *books* as the grammatical object here exerts the most influence on *I*, *see* and *two* – the subject and predicate in the sentence and a numeral that defines it.

We average the influence values per column, which gives a measure of the average influence a token exerts on its sentence embedding. For each token, we store this influence value together with its part-of-speech tag, tagged with spaCy (Honnibal et al., 2020). For all LMs, we fit a linear regression model for the influence value as a dependent variable, with the following independent variables[3] with statsmodels (Seabold and Perktold, 2010):

- Token position, to investigate whether a position bias exists
- Token length (in characters), to see how different parts of speech and/or higher information content affect the influence
- Sequence length (number of tokens), to see how longer sequences affect influence values
- Construction type, to see whether paradigmatic and syntagmatic differences mediate these effects

Furthermore, we calculate the average influence for part-of-speech categories for all model/construction combinations.

5 Results

5.1 Regression analysis

Table 2 shows the linear models for each investigated language model, reporting the intercept and the regression coefficients for token position, word length, sentence length, the construction type, and the R^2 for the respective regression. For the construction type, imperative sentences form the baseline, while the other two types were included as categorical variables, which means that their corresponding results signify their relative impact compared to the influence values for the imperative sentences. The test statistics of all models' *F*-tests, as well as those of the *t*-tests for all values, were statistically highly significant ($p < 0.001$). We acknowledge that this significance might also be caused by the high number of data points available.

[3]Position, token length and sequence length were normalized to values in the range $[0; 1]$ before fitting the linear regression models.

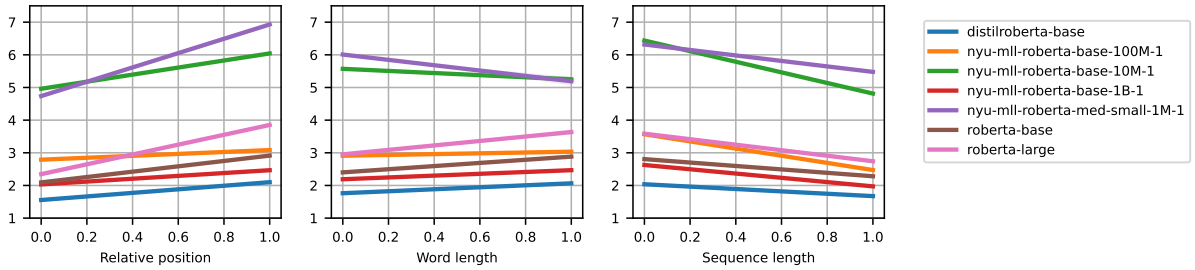

Figure 2: Regression lines for token position, word length and sentence length (calculated independently), with the influence value as the dependent variable

	roberta-med-small-1M-1	roberta-base-10M-1	roberta-base-100M-1	roberta-base-1B-1	roberta-base	roberta-large	distilroberta-base
(intercept)	4.871	5.367	3.188	2.115	2.012	2.186	1.478
token position	2.351	1.225	0.373	0.557	1.009	1.818	0.675
word length	1.075	0.978	0.538	0.834	1.287	2.129	0.883
sentence length	-0.932	-1.959	-1.210	-0.712	-0.550	-0.909	-0.402
cxn = wh-question	0.158	0.847	0.263	0.216	0.057	0.123	0.122
cxn = transitive	0.206	0.582	0.208	0.109	0.090	0.190	0.090
R^2	0.372	0.506	0.388	0.264	0.373	0.409	0.428

Table 2: Summary of parameter estimates for the logistic regression models across RoBERTa models of different size (column order roughly corresponds to model size). The response variable corresponds to the calculated influence values per word. The baseline for the categorical variable "construction" (*cxn*) are the imperative sentences included in the data set.

Across all models and constructions, the **token position** has a positive effect on the influence value. The same can be said for **word length**. This means that, for the current experiment, small and large LMs have a clear and systematic preference for putting more weight on sequence endings and longer words. This effect, to the best of our knowledge, has not been described before. Interestingly, for the two smallest models, the regression coefficient is larger for token position than for word length, a relation that is reversed for all other models trained on more data. **Sentence length** has a constantly negative effect on the influence value – longer sequences thus lower the influence values of their contained tokens. The effects of **construction type** on the influence value are generally positive, which points towards individual words' influence being higher in transitive sentences and wh-questions, when compared to the baseline (imperatives). These tendencies with regard to constructions are stable across all models.

Figure 2 shows model-wise regression lines only incorporating token position *or* word length *or* sentence length. The two models trained with the least amount of data (1M and 10M tokens) have the highest absolute influence values. This is also reflected in the comparatively high intercepts reported in Table 2. The other models' values are roughly equivalent. Interestingly, word length alone has a slightly negative effect for the two smallest models (1M and 10M). When incorporating all variables in the regression model, as in Table 2, this tendency is reversed. It is plausible that other interaction effects exist between the independent variables, which further underlines the importance of accounting for all of them in the regression model.

Finally, the R^2-values as goodness-of-fit measures exhibit considerable variation between the models. Training data and model size appear to play a certain role, as roberta-base-10M, the model with the least amount of data learned with a non-reduced architecture, features the highest R^2-value. However, the 10M model also features (by far) the largest regression coefficients for the construction types. Overall, the values range from 26.4% for the roberta-base-1B model to 50.6% for the 10M model. This shows that, although position, word length, sentence length and construction type cannot function as the sole predictors of influence values, they are influential variables under certain circumstances, possibly mediated by architectural factors beyond the scope of the present analysis.

Crucially, there is no linear relationship between the amount of training data, the model size and the goodness-of-fit, hinting towards an interconnectedness and/or more or less "fitting" combinations of internal factors like number of layers or attention heads, and model size.

5.2 Impact of construction type on part-of-speech influence

General overview In addition to the regression analysis, we calculated the average influence values on the part-of-speech level. These values are shown in Tables 3, 4 and 5, with the top three/four highest values set apart in bold face. Across all construction-model combinations, nouns, proper nouns and punctuation symbols are consistently influential.[4] For the models trained on more data, verbs are quite influential as well (at least for transitive sentences and wh-questions). For wh-questions and imperatives, also numerals are sporadically in the top three. Overall, the three different construction types show a similar, but still variable picture. Regarding model size, it is interesting to note that the smallest and the largest models (and distilroberta-base) tend to have their own, fairly stable rankings, whereas the most outliers occur in the medium-sized models.

Wh-questions The data for the wh-question in Table 3 is the least straightforwardly interpretable – only proper nouns are consistently influential. Apart from that, a contrast between the smallest/largest models and the medium-sized ones is noticeable. The smallest/largest models feature nouns and punctuation marks as most influential, whereas the medium models show larger influence values for numerals, verbs, and also (once) for adverbs and auxiliaries.

Transitive sentences For the transitive sentences (Table 4), nouns and punctuation marks are constantly among the most influential parts of speech. Here, a division can be drawn between the smaller and larger models. Whereas smaller models focus more on proper nouns, the larger models feature high influence values for verbs. For the roberta-base-1B model, one outlier is the auxiliary tag already found in the wh-question data.

[4]The X tag for unknown part of speech was consistently strong as well. A closer inspection of the dataset yields that the respective tokens are family-internal onomatopoeia or similar phonetic descriptions, which most probably are absent from the training data and thus *qua definitionem* more influential. Consequently, they were not set apart in bold face.

Imperatives The data for imperative sentences in Table 5 once more features punctuation and nouns as the most influential. Proper nouns are also highly influential, except for the two medium-sized (100M and 1B) models, where pronouns and auxiliaries also play a role.

6 Discussion

The present analysis in section 5.1 has shown that token position, word length and sequence length strongly affect sequence embeddings in terms of the influence of their lexical elements (viz. tokens). The general effect of token position on token influence is positive and stable across seven different RoBERTa models. All models exhibit a sequence-ending bias for the influence values. However, the effects exhibit variable strength. One reason for this could lie in the model size (training data, hyperparameters and model internals) – the effects (and the absolute influence values) are higher for smaller models but also for the largest model.

Apart from the token position, word length also has a constantly positive effect on the influence value. Longer words are thus more influential. The Zipfian law (Zipf, 1935) posits an inverse relationship between word frequency and length. As the most frequent words tend to be function words, the positive effect of word length on the influence value could also hint towards the higher informational content of longer words. Piantadosi et al. (2011) find a high correlation between word length and informational content for words in English, Swedish and German.

The negative impact of the sentence length could be caused by more tokens having to "share" the informational content of the whole sequence, which is then divided between all of them. The positive influence of the non-imperative sentences allows a ranking of construction type influence, where the effect is the weakest for imperatives, stronger for transitives and the strongest for wh-questions. From a linguistic point-of-view, the reasons for this remain elusive. Wh-questions and imperatives have more syntactically and lexically-fixed constructional schemas than transitive sentences. There is a possible connection between the functional aspects of imperatives and their reduced influence values, because they usually trigger real-world actions. In contrast, the information-driven functions of (information-demanding) questions and (information-conveying) regular transitive sen-

tences could impact their tokens' influence values positively.

When directly comparing the parts of speech, construction types do not seem to hold much explanatory value either. There is no systematic variation between their preferred parts of speech. In contrast, nouns, proper nouns, numerals and punctuation marks are consistently important across the types. Outliers are sporadic, only the higher influence of verbs in transitive sentences embedded through models with more training data is somewhat systematic. Importantly, including positional information is an active research topic in contemporary NLP (see Dufter et al., 2022 for a survey). The present results suggest that the token influence patterns already encode positional information, although transformers are theoretically invariant to the reordering of tokens in a sequence.

Comparing the data from a model-oriented perspective yields interesting, although ambiguous and inconclusive results. The smallest models (in terms of architecture and training data), as well as the largest models (in terms of training data) stabilize in different ways with regard to their most influential parts of speech. The medium-sized models (in terms of training data) exhibit more variation, focus on more exotic parts of speech and the corresponding linear regressions have a somewhat lower goodness-of-fit as well as lower overall regression coefficients. Crucially, it seems that for a stable and predictable functionality, a certain match between model size in terms of internal architecture (hidden layers, attention heads, etc.) is needed. Small data needs smaller models, and large data needs larger models. If these factors do not match, the representations become brittle and potentially less useful for downstream tasks. The concrete make-up of such matching combinations still needs more empirical scrutiny. For example, the model with the highest R^2, roberta-base-10M, also features the highest regression coefficients for the construction types. This relationship does not stabilize across the other model-data combinations, with no discernible reasons identifiable from the present analysis.

Also, as further empirical results show that the processing in LMs mirrors traditional NLP pipelines along the layers of linguistic processing (Tenney et al., 2019), the value of LMs for studies of linguistic processing has been put to question (Linzen and Baroni, 2021; Warstadt and Bowman, 2022). Pannitto and Herbelot (2022) argue that neural networks should also be used to investigate usage-based theories of language. The present study has added to this emergent field by showing that findings from usage-based linguistics on the importance of sequence order to language use are indeed mirrored in transformer-based LMs. However, the construction-level effects proposed in linguistic literature could not be completely verified. This might be due to the very different nature of language acquisition in humans and the training procedure in ANNs. Training only mirrors the frequency-driven aspect of usage-based linguistics. Other aspects like embodied cognition or the functional dimension of language, which can also be linked to construction types (e.g. in Cameron-Faulkner and Hickey (2011)), are missing. Remarkably, function words are not as influential as lexical words. Their structural predictability could be an influence factor in this case. Constructions are usually conceptualized as structures with open slots. Here, paradigmatic variation is much higher for lexical words, which are also more influential for models. However, the great amount of variation suggests that not all LMs learn the exact same structures, with inadequate data/model matchings leading to more brittle representations. Dąbrowska (2012) argues that the grammatical systems of adult speakers do not completely align with each other – they are only similar enough to enable effective communication. Judging from our results, the grammatical systems in language also feature different sensitivities to factors like word length or sequence length. This could point to learning with fitting parameter combinations being more human-adequate, as the linguistic and architectural effects on LMs are gradient in nature (a feature they share with human language processing and usage). Most importantly, this analysis has shown that the trade-offs between data size, model internals, and stable performance deserve further recognition and investigation, because mismatched combinations may lead to unstable or brittle representations.

7 Conclusion

Our investigation shed light on the functionality of LMs from a usage-based perspective, and has shown that concepts from usage-based linguistics, like entrenchment, can be used fruitfully in the analysis of such LMs. We discovered that frequency-driven factors, as well as information weight, play a significant role in these models' encodings. No-

tably, the models exhibit a bias towards the ends of sequences, with the influence of tokens positively correlated with their length and information-rich parts of speech, such as nouns. However, these effects weaken in longer sequences. The high variation across our statistical models' R^2-values hints at additional factors beyond entrenchment being at play when determining token influence on sequence embeddings. Still, our findings suggest that human learners and artificial learners share similarities, as both processes are influenced by frequency and information effects. Significant differences in influence values between construction types indicate a need for further research to interpret these differences linguistically. Additionally, our study explored the similarities and differences between models trained with varying amounts of data. While general effects remain similar, there is increased volatility, especially in preferred parts of speech, with shrinking data size. A non-linear relationship between the amount of training data, model architecture, and effect sizes/goodness-of-fit was observed. This highlights the need for deeper investigations into the optimal combinations of data and other hyperparameters.

Limitations

The present study is limited by the availability of models with different, yet comparable (e.g. in terms of training data or traceable stepwise adjustment) training regimens. More empirical results with regard to data size and model internals, investigated in a systematic and controlled way, are clearly needed. Furthermore, it would also be interesting to additionally look into the perturbation patterns for different layers in LMs, which could further illuminate the ways in which structural sensitivity mirrors the levels of human linguistic processing.

Acknowledgements

We acknowledge financial support by the project "SAIL: SustAInable Life-cycle of Intelligent Socio-Technical Systems" (Grant ID NW21-059A), which is funded by the program "Netzwerke 2021" of the Ministry of Culture and Science of the State of Northrhine Westphalia, Germany.

References

Jennifer E. Arnold, Anthony Losongco, Thomas Wasow, and Ryan Ginstrom. 2000. Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, 76(1):28–55.

Otto Behaghel. 1930. Von deutscher Wortstellung. *Zeitschrift für Deutschkunde*, 44:81–89.

Joan Bybee and Sandra Thompson. 1997. Three Frequency Effects in Syntax. *Annual Meeting of the Berkeley Linguistics Society*, 23(1):378.

Thea Cameron-Faulkner and Tina Hickey. 2011. Form and function in Irish child directed speech. *Cognitive Linguistics*, 22(3):569–594.

Thea Cameron-Faulkner, Elena Lieven, and Michael Tomasello. 2003. A construction based analysis of child directed speech. *Cognitive Science*, 27(6):843–873.

Noam Chomsky. 1957. *Syntactic Structures*. Mouton Publishers.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. Number 11 in Massachusetts Institute of Technology. Research Laboratory of Electronics. Special Technical Report. MIT Press, Cambridge, Massachusetts.

Ewa Dąbrowska. 2012. Different speakers, different grammars: Individual differences in native language attainment. *Linguistic Approaches to Bilingualism*, 2(3):219–253.

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*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Holger Diessel. 2019. Usage-based construction grammar. In Ewa Dąbrowska and Dagmar Divjak, editors, *Cognitive Linguistics - A Survey of Linguistic Subfields*, pages 50–80. De Gruyter.

Philipp Dufter, Martin Schmitt, and Hinrich Schütze. 2022. Position Information in Transformers: An Overview. *Computational Linguistics*, 48(3):733–763.

Charles J. Fillmore. 1988. The Mechanisms of "Construction Grammar". *Annual Meeting of the Berkeley Linguistics Society*, 14:35.

Adele E. Goldberg. 2003. Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences*, 7(5):219–224.

Zhi Hong, Aswathy Ajith, Gregory Pauloski, Eamon Duede, Carl Malamud, Roger Magoulas, Kyle Chard, and Ian Foster. 2022. ScholarBERT: Bigger is Not Always Better.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in python.

Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. BabyBERTa: Learning More Grammar With Small-Scale Child-Directed Language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646, Online. Association for Computational Linguistics.

Ronald W. Langacker. 1987. *Foundations of Cognitive Grammar. Vol. 1: Theoretical Prerequisites*, volume 1. Stanford University Press, Stanford, California.

Bai Li, Zining Zhu, Guillaume Thomas, Frank Rudzicz, and Yang Xu. 2022. Neural reality of argument structure constructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7410–7423, Dublin, Ireland. Association for Computational Linguistics.

Tal Linzen and Marco Baroni. 2021. Syntactic Structure from Deep Learning. *Annual Review of Linguistics*, 7(1):195–212.

Shusen Liu, Zhimin Li, Tao Li, Vivek Srikumar, Valerio Pascucci, and Peer-Timo Bremer. 2019a. NLIZE: A Perturbation-Driven Visual Interrogation Tool for Analyzing and Interpreting Natural Language Inference Models. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):651–660.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A Robustly Optimized BERT Pretraining Approach.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*, 3 edition. Lawrence Erlbaum Associates, Mahwah, NJ.

Brian MacWhinney. 2018. MOR Manual.

Benjamin Newman, Kai-Siang Ang, Julia Gong, and John Hewitt. 2021. Refining Targeted Syntactic Evaluation of Language Models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3710–3723, Online. Association for Computational Linguistics.

Ludovica Pannitto and Aurelie Herbelot. 2022. Can Recurrent Neural Networks Validate Usage-Based Theories of Grammar Acquisition? *Frontiers in Psychology*, 13:741321.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

Randolph Quirk, editor. 1972. *A Grammar of Contemporary English*. Longman, London.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter.

Hans-Jörg Schmid. 2015. A blueprint of the Entrenchment-and-Conventionalization Model. *Yearbook of the German Cognitive Linguistics Association*, 3(1).

Skipper Seabold and Josef Perktold. 2010. Statsmodels: Econometric and Statistical Modeling with Python. In *Python in Science Conference*, pages 92–96, Austin, Texas.

Ekaterina Taktasheva, Vladislav Mikhailov, and Ekaterina Artemova. 2021. Shaking Syntactic Trees on the Sesame Street: Multilingual Probing with Controllable Perturbations. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 191–210, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Harish Tayyar Madabushi, Laurence Romain, Dagmar Divjak, and Petar Milin. 2020. CxGBERT: BERT meets Construction Grammar. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4020–4032, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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. Association for Computational Linguistics.

Michael Tomasello. 2003. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.

Yu-Hsiang Tseng, Cing-Fang Shih, Pin-Er Chen, Hsin-Yu Chou, Mao-Chang Ku, and Shu-Kai Hsieh. 2022. CxLM: A construction and context-aware language model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6361–6369, Marseille, France. European Language Resources Association.

Alex Warstadt and Samuel R. Bowman. 2022. What Artificial Neural Networks Can Tell Us About Human Language Acquisition.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020a. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Alex Warstadt, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R. Bowman. 2020b. Learning Which Features Matter: RoBERTa Acquires a Preference for Linguistic Generalizations (Eventually). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 217–235, Online. Association for Computational Linguistics.

Leonie Weissweiler, Taiqi He, Naoki Otani, David R. Mortensen, Lori Levin, and Hinrich Schütze. 2023. Construction Grammar Provides Unique Insight into Neural Language Models. In *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)*, pages 85–95.

Leonie Weissweiler, Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze. 2022. The better your syntax, the better your semantics? Probing pretrained language models for the English comparative correlative. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10859–10882, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick Von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. 2020. Perturbed Masking: Parameter-free Probing for Analyzing and Interpreting BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4166–4176, Online. Association for Computational Linguistics.

Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. 2021. When Do You Need Billions of Words of Pretraining Data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1112–1125, Online. Association for Computational Linguistics.

George K. Zipf. 1935. *The Psycho-Biology of Language*. Houghton Mifflin, Boston.

A Construction retrieval patterns

The corpus files in CHILDES come with morphological annotation in form of the CHILDES tag set (MacWhinney, 2018, 8). To match the constructions, the following (high-level) patterns were considered and implemented via regular expressions:

- **Imperatives:** [adverb] + verb/modal + [quantifier] + [adverb] + [determiner] + noun/pronoun
- **Wh-questions:** [preposition] + interrogative pronoun/interrogative determiner/relative pronoun/conjunction + ?
- **Transitive sentences:** personal pronoun/subject pronoun/indefinite pronoun/noun/proper noun/demonstrative determiner + [modal/auxiliary] + [adverb] + [quantifier] + personal pronoun/subject pronoun/indefinite pronoun/noun/proper noun/demonstrative determiner

NB: The quality and consistency of morphological annotations varies considerably between the CHILDES data sets. Consequently, some part-of-speech tags had to be included for the pattern matching that make the constructional schemas appear intuitively wrong. Furthermore, questions were pre-annotated with question marks.

B Full influence values for all construction types

part of speech	roberta-med-small-1M	roberta-base-10M	roberta-base-100M	roberta-base-1B	roberta-base	roberta-large	distilroberta-base
ADJ	5.4	5.78	3	2.47	2.69	3.62	1.96
ADP	5.28	5.09	2.9	2.32	2.42	3.12	1.85
ADV	5.66	5.64	**3.24**	2.47	2.68	3.54	1.97
AUX	5.45	5.62	3.06	**2.81**	2.52	2.88	1.8
CCONJ	4.47	4.25	2.07	1.05	1.26	1.39	0.94
DET	4.56	4.97	2.76	2.1	2.06	2.45	1.58
INTJ	4.13	4.35	2.42	1.22	1.45	1.44	1.19
NOUN	**5.71**	5.77	3	2.4	**2.76**	**3.66**	**2.03**
NUM	5.7	**5.88**	**3.17**	**2.56**	2.66	**3.74**	1.95
PART	5.47	4.65	2.63	2.09	2.21	2.87	1.65
PRON	4.64	5.44	2.82	1.98	2.08	2.26	1.6
PROPN	**6.37**	**6.57**	**3.46**	**2.56**	**2.76**	**3.73**	**2.14**
PUNCT	**8.27**	**6.74**	3.1	2.46	2.73	3.56	**2.07**
SCONJ	4.12	4.89	2.63	1.32	1.76	1.41	1.42
VERB	5.54	5.85	3.11	**2.6**	**2.77**	3.6	1.98
X	7.15	6.96	3.36	3.09	3.16	3.92	2.29
(mean)	5.50	5.53	2.92	2.22	2.37	2.95	1.78

Table 3: Average influence values for wh-questions

part of speech	roberta-med-small-1M	roberta-base-10M	roberta-base-100M	roberta-base-1B	roberta-base	roberta-large	distilroberta-base
ADJ	5.93	5.36	2.79	2.26	2.57	3.35	1.9
ADP	5.59	4.64	2.65	2.04	2.32	2.99	1.79
ADV	5.85	5.25	2.78	2.25	2.47	3.15	1.83
AUX	5.71	5.24	2.77	**2.31**	2.28	2.89	1.7
CCONJ	4.62	4.48	2.4	1.23	1.39	1.55	1.09
DET	5.34	4.68	2.63	2.02	2.12	2.64	1.6
INTJ	4.55	4.49	2.4	1.36	1.53	1.57	1.23
NOUN	**6.38**	**5.69**	**2.91**	**2.3**	**2.8**	**3.54**	**2.03**
NUM	5.86	5.3	2.69	2.26	2.52	3.18	1.81
PART	5.33	4.18	2.39	1.85	2.09	2.49	1.65
PRON	5.32	5.07	2.68	1.96	2.24	2.49	1.56
PROPN	**6.6**	**6.02**	**3.09**	2.23	2.61	3.2	1.91
PUNCT	**6.44**	**6.13**	**3.09**	2.16	**2.66**	**3.41**	**1.92**
SCONJ	4.52	4.36	2.44	1.47	1.67	1.63	1.29
VERB	5.59	5.36	2.92	**2.46**	**2.65**	**3.42**	**1.92**
X	7.19	6.1	2.98	2.41	2.88	3.6	2.11
(mean)	5.68	5.15	2.73	2.04	2.30	2.82	1.71

Table 4: Average influence values for transitive sentences

part of speech	roberta-med-small-1M	roberta-base-10M	roberta-base-100M	roberta-base-1B	roberta-base	roberta-large	distilroberta-base
ADJ	5.87	5.25	2.93	2.32	2.75	3.44	1.93
ADP	5.74	4.86	3.03	2.36	2.55	3.36	1.87
ADV	6.03	5.4	**3.05**	2.35	2.64	3.36	1.88
AUX	5.72	5.26	2.69	2.08	2.24	2.81	1.68
CCONJ	5.16	4.74	2.65	1.83	1.92	2.44	1.35
DET	5.52	4.8	2.89	2.3	2.33	2.85	1.71
INTJ	4.99	4.76	2.5	1.55	1.73	1.89	1.37
NOUN	**6.73**	**5.71**	**3.06**	**2.43**	**3.09**	**3.89**	**2.16**
NUM	5.96	5.45	2.92	**2.47**	2.73	**3.57**	1.9
PART	5.44	4.33	2.61	1.94	2.12	2.48	1.54
PRON	5.67	5.23	2.91	**2.43**	2.56	3.22	1.78
PROPN	**6.89**	**5.96**	2.94	2.34	**2.9**	**3.66**	**2.05**
PUNCT	**6.75**	**6.19**	**3.22**	2.29	**2.76**	**3.57**	**1.98**
SCONJ	5.26	4.77	2.75	2.17	2.48	3.07	1.75
VERB	4.91	5.15	2.95	2.17	2.29	2.57	1.67
X	7.6	6.26	3.45	2.79	3.15	4.03	2.36
(mean)	5.89	5.26	2.91	2.24	2.52	3.14	1.81

Table 5: Average influence values for imperatives