Vocabulary Shapes Cross-Lingual Variation of Word-Order Learnability in Language Models

Anonymous ACL submission

Abstract

Why do some languages like Czech permit free word order, while others like English do not? We address this question by pretraining transformer language models on a spectrum of synthetic word-order variants of natural languages. We observe that greater word-order irregularity consistently raises model surprisal, indicating reduced learnability. Sentence reversal, however, affects learnability only weakly. A coarse distinction of free- (e.g., Czech and Finnish) and fixed-word-order languages (e.g., English and French) does not explain cross-lingual variation. Instead, the structure of the word and subword vocabulary strongly predicts learnability. Overall, vocabulary structure emerges as a key driver of computational word-order learnability across languages.

Code repository (anonymized for review)

1 Introduction

Human languages have emerged over millennia through dynamics shaped by communicative and cognitive constraints (Zipf, 1935; Piantadosi et al., 2012; Hawkins, 2014; Futrell et al., 2020; Hahn and Xu, 2022; Clark et al., 2023). Yet, within those universally shared bounds, languages exhibit a striking typological diversity, varying in morphological complexity and preferred word orders, for example. Languages, in all their diversity, are not equally complex in every aspect (Croft, 2002; Sampson et al., 2009; Koplenig et al., 2023). This raises a central question: Are all languages equally hard to learn? And if not, why?

One dimension of linguistic diversity is word-order flexibility—the degree to which words in a sentence can be reordered without changing its meaning, except for emphasis. In Czech, for instance, case marking determines the grammatical role of nouns in a sentence, allowing constituent order to vary relatively freely. In the sentence "Robot maluje kočku." (*The robot paints the cat.*), any of the six permutations of subject (robot), verb (maluje), and object (kočku, the accusative case of kočka) is grammatically acceptable and conveys the same core meaning. In English, by contrast, the sentence "The robot paints the cat." cannot be reordered without changing its meaning or rendering it ungrammatical.

Research questions The Czech–English example illustrates a general typological pattern: Languages with relatively free word order (like Czech) tend to encode syntactic relations through morphology while with relatively fixed word order (like English) rely on word position instead. This contrast motivates two questions: First, whether learnability is sensitive to the degree of word-order flexibility; and second, why some languages are more robust to free word order than others.

Synthetic languages In natural languages, typological features are often strongly correlated (Greenberg, 1990). Synthetic-language experiments aim to solve this problem by perturbing a natural language along a single dimension, for example altering word order while preserving vocabulary and content (Kallini et al., 2024; Xu et al., 2025; Yang et al., 2025). However, prior work has faced two limitations: First, word order flexibility and morphological complexity are mostly studied in isolation, although these two factors are clearly connected (Bisazza et al., 2021; Nijs et al., 2025; Liu et al., 2025). Perturbations experiments commonly operate at the subword level, which often breaks up lexical units. For example, a subword tokenizer might split the Czech word *maluje* into *ma* and *luje*, which yields linguistically implausible sequences when shuffled. Thus, word order and morphology are perturbed simultaneously. Second, the use of disparate shuffling methods with discrete parameters makes it difficult to compare results and limits control over perturbation strength. Due to

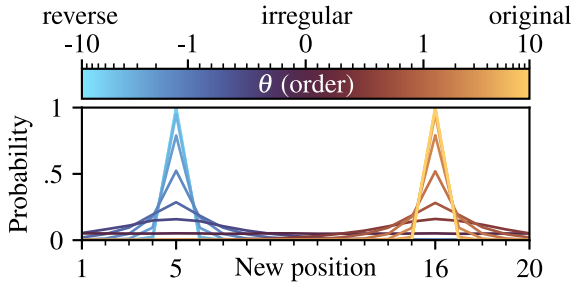

Figure 1: We create a spectrum of synthetic language variants by deterministically permuting words within each sentence. For each sentence length, a permutation is sampled from the Mallows permutation model, where the order parameter θ controls preference for the original word order. As an example, we show the probability distribution of a word originally at position 16 in a 20-word sentence.

these limitations, the interplay of word-order flexibility, morphological complexity, and tokenization in shaping computational learnability remains an open question (Arnett and Bergen, 2025; Poelman et al., 2025).

Approach and contributions To overcome these limitations in answering our two research questions, we design a controlled cross-lingual perturbation experiment. We create a continuous spectrum of synthetic word-order variants for ten European languages by deterministically shuffling at the word level. Our approach uses the Mallows permutation model, which provides a single continuous parameter, the *order* θ, that controls the regularity of word order. This parameter can be interpreted as a preference for the original word order: Large positive values correspond to the original order; small positive values yield local shuffling; at $\theta = 0$, the order is irregular, such that every word order is equally likely; and negative θ corresponds to aversion to the original order, up to sentence reversal, see Fig. 1. Crucially, by deterministically shuffling whole words rather than subwords, our method preserves the model-independent global text entropy, vocabulary, and morphology of the original sentences, ensuring that the language variants differ only in terms of word-order regularity.

Our experiments reveal two key findings: First, by shuffling on the word level rather than subwords, we confirm that language model surprisal increases with more irregular word order, yet it is largely insensitive to sentence reversal. Second, categorical word-order typology fails to account for language-specific differences, as word-order flexibility is rather a gradient. Instead, vocabulary statistics—Zipf-based coverage metrics, sentence length, and simple proxies for morphological complexity—explain well how robust a language is to free word order in terms of learnability.

2 Language learnability

Thousands of natural languages exist worldwide (Hammarström et al., 2025), displaying a wide variety in structural patterns. Here, we are interested in the way these characteristic features influence how difficult a language is to learn for humans and computational models. This section reviews the relation of typological variation to learnability.

2.1 Language variation

Natural languages evolve under cognitive and communicative constraints shared by all humans, including limits on information density, redundancy, and processing load (Zipf, 1935; Piantadosi et al., 2012; Hawkins, 2014; Hahn and Xu, 2022; Futrell et al., 2020; Clark et al., 2023). Within the vast space of symbolic communication systems, natural languages are a small subset shaped by typological correlations (Greenberg, 1990). Yet, despite these common forces, they exhibit a striking structural variety.

One prominent example of variation is *word-order flexibility*. The order of subject (S), verb (V), and object (O) is far from uniform across languages: Although the orders SVO or SOV dominate globally (Dryer and Haspelmath, 2024), many languages—such as those from the Slavic and Uralic families—permit comparatively free constituent order, relying on fusional or agglutinative morphology to encode syntactic relations (Ponti et al., 2019; Liu et al., 2025; Nijs et al., 2025; Svenonius, 2025). Certain registers, e.g., poetic Latin, even allow nearly unconstrained order (Sampson, 2009).

Word-order structure has been linked to principles such as entropy minimization (Franco-Sánchez et al., 2024) or uniform information density (Clark et al., 2023). Because many factors, especially complex morphology, are intricately connected with word-order flexibility, our goal is to disentangle their contributions to language learnability.

2.2 Computational learnability

Not all languages are equally complex (Sampson et al., 2009; Koplenig et al., 2023), but it remains

unclear whether language models can learn all languages equally well—be they artificial or natural—or whether current architectures systematically favor certain linguistic features.

In this article, we focus on computational learnability, i.e., how well a model captures the probability distribution of a language, rather than human learnability. Language models are useful in this context because they offer controlled, large-scale experimental setups impossible with human subjects for testing linguistic hypotheses (Piantadosi, 2024; Futrell and Mahowald, 2025).

Empirical studies indicate that languages differ in how easily models acquire them. Complex inflectional morphology might make languages more difficult (Cotterell et al., 2018), although subsequent work found simpler statistics like vocabulary size and length in characters to be more predictive of model performance than linguistic factors (Mielke et al., 2019).

In models, tokenization further complicates matters. A performance gap between agglutinative and fusional languages appears to be driven by encoding efficiency rather than morphology itself (Arnett and Bergen, 2025). However, tokenization properties—including productivity, idiosyncrasy, and fertility—are in turn closely tied to morphology (Gutierrez-Vasques et al., 2023). Measuring morphological features and linking them to learnability is further complicated by the fact that many typological features are more accurately described as gradients than as discrete classes (Levshina et al., 2023; Baylor et al., 2024; Poelman et al., 2025).

3 Methodology

One way to disentangle typology and learnability is to create *synthetic language variants* that systematically alter a single typological dimension while keeping the others intact. When based on natural languages, these variants preserve the complexity and irregularity of the original languages, but allow targeted manipulations. This idea has been used to explore phenomena like non-concatenative morphology (Haley and Wilson, 2021) or translating free-word-order variants of fixed-word-order languages (Bisazza et al., 2021). We build on this approach by generating synthetic word-order variants to isolate how word order and vocabulary shape learnability.

3.1 Synthetic word order

Experiments using word-order perturbations probe how sequence structure affects language models. Subword-level shuffling has shown that perturbations harm transformer performance, e.g., Kallini et al. (2024). However, random permutations increase the entropy of the text, complicating the interpretation of model surprisal. Deterministic permutations avoid raising the model-independent entropy by using fixed permutations for each sentence length (Clark et al., 2023; Someya et al., 2025).

Recent cross-lingual studies of computational learnability arrive at mixed results. Ziv et al. (2025) found no consistent preference for natural over artificial languages, while Yang et al. (2025) report moderate inductive bias in favor of natural languages but invariance to violations of certain typological correlations. In contrast, targeted manipulations of specific typological correlations indicate a weak learning bias against those variants (Xu et al., 2025; El-Naggar et al., 2025b).

These studies highlight the value of word-order perturbation for probing learnability, but several limitations recur. First, *perturbations of subwords* split words at inconsistent places that are not linguistically meaningful, which alters both syntax and morphology simultaneously (Beinborn and Pinter, 2023; Di Marco and Fraser, 2024). Second, *disparate shuffling methods* rely on discrete parameters, hindering comparability and control over the degree of perturbation. Third, *narrow language selections*, typically restricted to English, leave cross-lingual variation underexplored.

3.2 Deterministic shuffling

We address these limitations of prior perturbation studies through deterministic *word-level shuffling* with a single continuous order parameter θ, applied to a multilingual parallel corpus. This design preserves morphology and keeps model-independent global entropy practically constant, enabling systematic study of how vocabulary and word-order typology interact in determining learnability.

Our approach Intuitively, the desired control parameter *order* θ encodes a preference for the original word order of a given sentence in the corpus. By varying θ, we cover the whole spectrum of word-order regularity, ranging from the original order ($\theta \to \infty$), through locally shuffled ($\theta > 0$), to completely irregular ($\theta = 0$), to local shuffling of the reverse order ($\theta < 0$) and full sentence reversal

($\theta \to -\infty$), see Fig. 1.

For example, the sentence *the robot paints the cat* has five words, so we denote the original order as $\pi_0 = (1, 2, 3, 4, 5)$. At $\theta = 1$, we might sample a locally shuffled permutation $\pi = (2, 1, 3, 5, 4)$ corresponding to *robot the paints cat the*. At $\theta = 0$, all permutations π are equally likely. At $\theta = -7$, the sequence most likely reverses to $\pi = (5, 4, 3, 2, 1)$, i.e., *cat the paints robot the*.

Formal model We use the *Mallows ϕ model* (Mallows, 1957), which offers exactly the desired parameter, as the key element of our design. The Mallows model assigns the probability of a permutation $\pi \in \mathfrak{S}_n$ based on the distance d from the original word order $\pi_0 = (1, 2, \ldots, n)$ as

$$\mathbb{P}_{\theta,\pi_0,d}(\pi) = \frac{1}{Z(\theta, d)} \mathrm{e}^{-\theta d(\pi,\pi_0)} \qquad (1)$$

with the order parameter θ and a normalization Z (Crispino et al., 2023). Here, the distance metric d is Kendall's τ (Kendall, 1938; Tang, 2019), which counts the minimum number of adjacent swaps to restore the original order π_0 from the permutation π. With Kendall's τ, the probability distribution is easy to sample from (Fligner and Verducci, 1986) and yields local shuffling for large $|\theta|$. Technical details of the Mallows model and an efficient sampling algorithm are given in Appendix A.

Implementation For each sentence length $n = 1, \ldots, 80$ in the corpus of a given language, we sample a single permutation $\pi^{(n)}$ from the Mallows model and apply it to all sentences of that length. This makes the transformation deterministic, ensuring that the minimum description length (or, equivalently, the model-independent entropy) of the text increases only marginally[1] due to the additional information contained in the n permutations (Clark et al., 2023; Someya et al., 2025).

4 Experimental setup

For our experiments, we generate variants of natural languages with perturbed word order and train identical language models on each variant. This section outlines the training corpus and languages, pre-processing, shuffling algorithm, model and evaluation metrics.

[1]The model-*dependent* entropy for a left-to-right prediction objective may still be sensitive to this nonlocal component, since it cannot know the sentence length in advance.

4.1 Data

Corpus We require a parallel training corpus that encompasses multiple languages with different typology, high quality, and uniform register from multiple speakers, ideally with longer sentences for which word order plays a significant role. The Europarl corpus of European parliamentary speeches meets these criteria (Koehn, 2005).

Language selection For interpretability and computational feasibility, we focus on ten out of the 21 languages in Europarl: five languages typically classified as fixed-word-order and five as free-word-order, ensuring variation across morphological type (analytic, fusional, agglutinative). Note that typological categories, including word-order flexibility, are often more aptly described as gradients rather than discrete classes (Levshina et al., 2023; Baylor et al., 2024). Our sample comprises French, Portuguese (Romance); English, Swedish, Danish (Germanic); Latvian (Baltic); Czech (Slavic); Hungarian, Estonian, Finnish (Finno-Ugric), see Appendix B for details.

Data preparation The definition of a word is convoluted (Haspelmath, 2023). We define a word pragmatically as an orthographic unit (whitespace-delimited) to preserve morphological integrity.

Preprocessing involves lowercasing all words and removing punctuation to eliminate positional cues from brackets, commas, quotation marks, etc., see Appendix C. For each language, we remove sentences longer than 80 words, and then split into training, validation, and test sets of 650 000, 5000, and 5000 sentences, respectively.

4.2 Model

We train a lightweight autoregressive language model from scratch with the PICOLM framework (Diehl Martinez et al., 2025), a transformer architecture similar to LLAMA models but designed for reproducible research with small language models. The data is tokenized using ByteLevel-BPE with vocabulary size $|V| = 16000$ unless varied. All hyperparameters are listed in Appendix D.

4.3 Evaluation

We quantify the learnability of a synthetic language variant with order θ via *model surprisal* S. Surprisal measures how unexpected the observed next subword w_i is to the model given the preceding con-

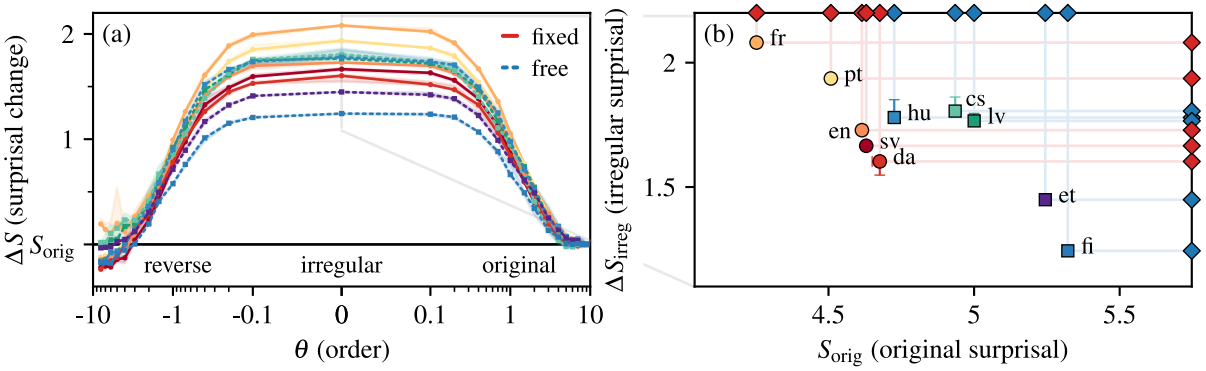

Figure 2: (a) Surprisal change ΔS due to word-order perturbations with order θ for each language (named in panel b). Color shades encode word order: fixed as solid-red and free as dashed-blue. (b) Zoom-in of surprisal change ΔS_{irreg} at irregular order $\theta = 0$ against the original surprisal S_{orig}. Red and blue markers on the axes indicate an overlap of free- and fixed-word-order languages in ΔS_{irreg} but not in S_{orig}. Transparent bands in panel (a) and error bars in (b) show the 25th to 75th percentile over seeds; the lines and points are the median seed, respectively.

text $w_{<i}$, i.e., the average negative log-probability

$$S(\theta) = \frac{1}{N} \sum_{i=1}^{N} -\log\big(p_\theta(w_i \mid w_{<i})\big), \qquad (2)$$

where p_θ is the model's predictive distribution and N is the total number of tokens in the sequence.[2] Each model is evaluated on a test set of the same language variant of order θ it was trained on.

From an information-theoretic perspective, surprisal is closely related to *entropy*—where entropy is the average over the Shannon information content (or surprisal) of each single outcome (Shannon, 1948; MacKay, 2019).

Entropy, and by extension the average surprisal, thus characterize compressibility (Schürmann and Grassberger, 1996): *Lower surprisal* means that the model has captured more of the sequence structure, reflecting *greater learnability* of that language variant. Since our shuffling method leaves global entropy essentially unchanged, any change in surprisal $S(\theta)$ relative to the original surprisal $S_{\text{orig}} = S(\theta \to \infty)$ of each language, $\Delta S(\theta) = S(\theta) - S_{\text{orig}}$, is due to the model's sensitivity to the word-order perturbations.

5 Word-order robustness

We now turn to how model surprisal varies across language variants for different orders θ, which governs the preference of a given language variant toward the original word order. Higher learnability means lower surprisal change ΔS relative to the unperturbed baseline S_{orig}.

Surprisal sensitivity First, we observe in Fig. 2 (a) that, across languages, language model surprisal increases with more irregular word order. The surprisal change ΔS is largest around the fully irregular word order ($\theta = 0$).

Furthermore, sentence reversal ($\theta < 0$) yields almost the same surprisal as the corresponding positive order perturbations ($\theta > 0$)[3]. This reflects the symmetry in θ of the Mallows model (Fligner and Verducci, 1986), which is largely preserved by the model surprisal, indicating that the models are not strongly sensitive to the typological correlations violated by reversal.

Cross-lingual differences Beyond the overall sensitivity to word-order perturbations observed above, the robustness to shuffling differs by language. Languages allowing freer word order (blue) substantially overlap in ΔS with languages that clearly prefer fixed word order (red), suggesting the former are, as a group, no more robust to perturbations, see Fig. 2 (a).

The distinction by free versus fixed word order alone is indeed insufficient: In panel (b), the two groups are clearly separated in baseline surprisal S_{orig}, yet they overlap in irregular surprisal

[2] For the full corpus, we calculate surprisal per subword token over non-overlapping batches due to finite context size.

[3] The magnitude of this effect is minor with a median surprisal asymmetry $\Delta S^{+/-}$ of 0.096 across θ, i.e., about 6 % of the surprisal change due to irregular word order. However, a Wilcoxon signed-rank test on paired differences $\Delta S^{+/-} = \Delta S^{+} - \Delta S^{-}$, aggregated per language, reveals a significant small asymmetry ($p = 0.0098$).

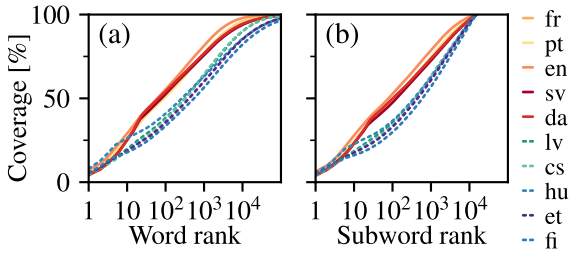

Figure 3: Percentage of (a) words and (b) subwords in the corpus accounted for by the most frequent vocabulary items. This coverage increase more slowly for languages with freer word order compared to languages with relatively fixed word order (shades of blue and red, respectively).

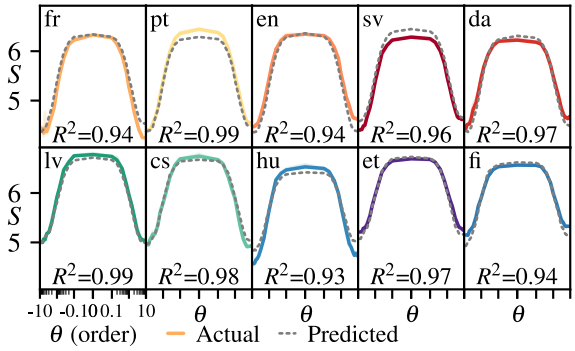

Figure 4: The absolute surprisal $S(\theta) = S_{\text{orig}} + \Delta S(\theta)$ per language modeled through a set of vocabulary statistics, encompassing coverage, sentence length, and proxies for morphological complexity. The predictions are cross-validated from leave-one-language-out: Each language is predicted solely on the basis of its own vocabulary statistics by a model trained on the surprisal of other languages and their predictors.

$\Delta S_{\text{irreg}} := \Delta S(\theta = 0)$. A Wilcoxon–Mann–Whitney test of the binary word-order flexibility on ΔS_{irreg} yields no significant difference between the groups at irregular word order ($p = 0.55$). Only the extremes—Romance (French, Portuguese) and Finnic (Finnish, Estonian)—are distinguished by both measures. This overlap suggests that factors beyond word-order flexibility drive cross-lingual variation.

6 The role of the vocabulary

Vocabulary structure, in the sense of word and subword frequencies, the relation of subwords to words, and sequence length, characterizes a language beyond word order. In fact, differences in Zipfian distributions (Piantadosi, 2014) relate closely to morphological complexity and word order (Liu et al., 2025). Vocabulary structure thus varies systematically between languages. Our aim is to derive latent structures from a set of simple metrics of vocabulary structure that explain cross-lingual variation in robustness to free word order.

6.1 Vocabulary metrics

One aspect of vocabulary structure is coverage $C(r)$: the proportion of the corpus accounted for by the r most frequent word or subword types. Coverage is the cumulative sum of the rank-frequency distribution described by Zipf's law (Zipf, 1949).

In Fig. 3 (a), word coverage clearly clusters languages into free word order (blues) and fixed word order (reds). Subword coverage in panel (b) preserves this separation after tokenization. This intrinsic typological grouping, which also captures the effect of tokenization on the vocabulary, renders vocabulary structure a strong candidate for explaining cross-lingual variation in surprisal.

This clustering suggests that coverage offers a more informative basis for predicting and thus explaining cross-lingual surprisal than a binary free/-fixed typology. To capture the essence of the coverage curves, we select four characteristics: word and subword coverage at rank 100, the integral of word coverage, and the similarity between word and subword coverage, defined in Appendix E.1. We complement this predictor set with other simple metrics of vocabulary structure: sentence length (average words and subwords) and proxies for morphological complexity (fertility, average word length, number of unique word types), see Appendix E.2.

6.2 Explaining word-order robustness

To identify latent structure in the predictors and assess their explanatory power for language-specific surprisal, we employ multivariate partial least squares (PLS) regression. PLS is well-suited for this setting of highly collinear predictors[4] and small sample size ($n = 10$ languages) with multivariate responses ($S(\theta)$ at 28 values of θ per language). PLS accomplishes dimensional reduction by creating latent components while retaining predictive power to explain the variance between predictors and response variables. Leave-one-language-out cross-validation identifies two components as the optimal number[5], with an overall predictive perfor-

[4] See correlation matrix in Appendix E.3.

[5] See Fig. 10 (a) and (b) in Appendix E.4 for details.

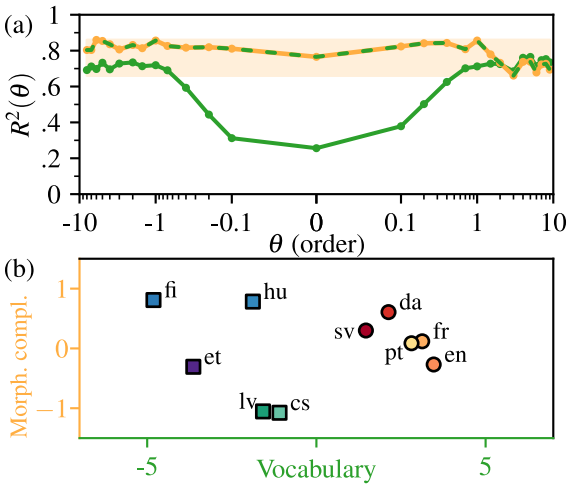

Figure 5: (a) The cross-validated explained variance per slice of θ of only the vocabulary component (green) with mean $\bar{R}^2 = 0.65$, ranging from 0.26 to 0.76 and of both components (green-yellow) with mean $\bar{R}^2 = 0.79$, ranging from 0.66 to 0.86. (b) PLS scores of the main component (vocabulary) and the secondary component (morphological complexity).

mance of $R^2 = 0.97$ variance explained.

Predictions from the cross-validated models capture the curve $S(\theta)$ closely and explain most of the variance per left-out language, ranging from $R^2 = 0.93$ for Hungarian to $R^2 = 0.99$ for Portuguese and Latvian, see Fig. 4.

Variance explained per slice of θ ranges from $R^2 = 0.66$ to 0.86 with mean $\bar{R}^2 = 0.79$, see Fig. 5 (a), demonstrating that the predictions are stable across various forms of word-order perturbations. The first component (vocabulary) alone explains original and reverse order with $\bar{R}^2 = 0.65$, ranging from 0.26 to 0.76. The second component (complex morphology: unique word types and word length) is therefore necessary to explain the regime of irregular word order.

Figure 5 (b) shows the learned latent structure. The primary component comprises coverage, and to a lesser extent sentence length and morphological complexity and structurally aligns equally across all θ. The secondary component reflects morphological complexity and is most associated with irregular order perturbations at small order $|\theta|$, see Fig. 10 (b) in Appendix E.4.

In summary, the vocabulary metrics explain surprisal $S(\theta)$ across languages and perturbation orders θ better than the binary free/fixed-word-order typology. While coverage explains the original and reverse order surprisal, complex morphology is a main factor of what makes a language more robust to shuffling.

6.3 Vocabulary size

Tokenization compresses the word vocabulary into a subword vocabulary and may influence cross-lingual differences in word-order robustness. We examine this by varying the vocabulary size $|V|$ for the original ($\theta = \infty$) and irregular word order ($\theta = 0$) condition, see Fig. 6.

The original surprisal S_{orig} begins to separate the free- and fixed-word-order languages above $|V| = 8000$, see panel (a), whereas the surprisal change at irregular word order ΔS_{irreg} in panel (b) converges between the languages at larger vocabulary sizes. This overlap stems from a downward trend or plateau of languages with rather fixed word order, while the other languages keep increasing up to $|V| = 64\,000$.

Panel (c) shows that this separation in S_{orig} coincides with clustering in subword coverage: Free-word-order languages make greater use of low-frequency subwords, consistent with the observation that the PLS component that explains the original surprisal is strongly associated with coverage.

7 Discussion

Our experiments show that higher word-order irregularity hinders language model learning across languages, but models remain largely insensitive to reversal. Cross-lingual variation is better predicted by vocabulary structure than by binary word-order flexibility. Our results clarify the factors that contribute to computational word-order learnability.

Relation to prior work The sensitivity to irregular word order reflects a *locality bias* (Choshen and Abend, 2019), and extends recent work on artificial-language learnability (Kallini et al., 2024; Xu et al., 2025; Yang et al., 2025; Kallini and Potts, 2025; El-Naggar et al., 2025b) to controlled word-level shuffling. By using a unified perturbation spectrum, our approach preserves morphological integrity and avoids confounds of disparate shuffling schemes in earlier work. Furthermore, our findings challenge claims that language models can learn all languages alike (Chomsky, 2023; Katzir, 2023; Leivada et al., 2025; Ziv et al., 2025).

Previous studies on sentence reversal found small and inconsistent differences in surprisal (Yang et al., 2025; Ziv et al., 2025). In contrast, our broader analysis across the continuous order

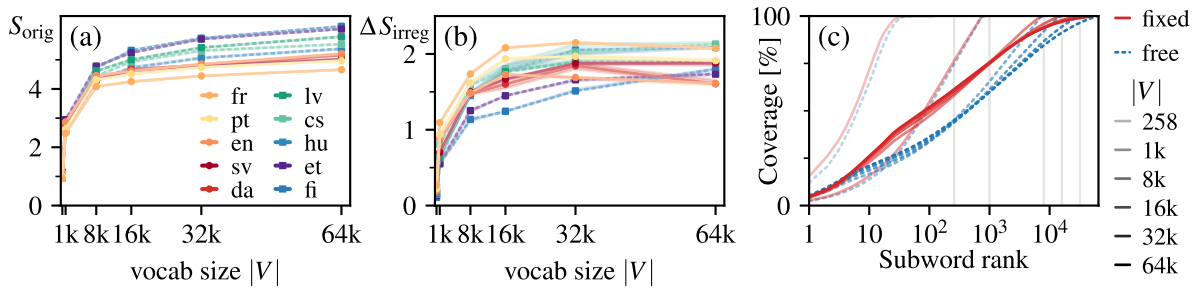

Figure 6: Effect of the vocabulary size on (a) the original surprisal S_{orig} and (b) the surprisal change at irregular word order ΔS_{irreg}. (c) Subword coverage per vocabulary size, grouped by free and fixed word order, where line transparency and the vertical gray lines encode the vocabulary size.

spectrum exhibits near symmetry with respect to reversal, slightly favoring reverse variants, matching a bias towards head-initial synthetic grammars El-Naggar et al. (2025a).

For *cross-lingual differences*, our results show in line with Mielke et al. (2019) that simpler vocabulary statistics and sentence length suffice as predictors, whereas Cotterell et al. (2018) emphasize the role of complex morphology. The apparent distinction between typology or simpler statistics as an explanation (Yang et al., 2025; Arnett and Bergen, 2025) is resolved if the vocabulary-based measures quantify aspects of typology—such as word-order robustness—more richly than coarse labels (Levshina et al., 2023; Baylor et al., 2024).

Architecture and mechanisms Several architectural factors may play into these results. First, the prediction objective of autoregressive transformer models limits the context for predicting the next token to previous tokens. Interestingly, since our deterministic shuffling can, in principle, be reversed if the sentence length is known, this shuffling introduces a nonlocal component to the language because the model does not know the sentence length ahead of time. Thus, the autoregressive nature of the models may underlie the general sensitivity to word-order perturbations across languages. Masked language models might therefore be less sensitive to irregular word order.

Second, larger vocabularies tend to reduce the irregular surprisal change for a subset of languages in our experiments. At large vocabulary sizes, the embedding parameters begin to outnumber the core model parameters. Possibly, languages that rely more on rare subwords and have high type diversity (see Appendix E.2), may be disadvantaged by limited model capacity (Tao et al., 2024).

Positional encodings also affect how shuffled input is represented, with distinct correlation patterns for words and subwords (Abdou et al., 2022). Future work should disentangle the architectural features of prediction objective, tokenization, and positional encoding.

8 Conclusion

We set out to understand what makes a language computationally difficult to learn for language models, using a spectrum of synthetic language variants with perturbed word order. Our experiments reveal three main findings: (1) Irregular word order decreases computational learnability, (2) but language models are largely insensitive to subtler violations of typological correlations introduced by sentence reversal; and (3) the robustness of a language to word-order perturbations is predicted better by vocabulary structure (Zipf-based coverage, sentence length, and morphological complexity) than by the coarse distinction into free and fixed word order. Morphological complexity proxies are most relevant for explaining robustness against strongly irregular word order.

These findings establish that simple vocabulary metrics provide a powerful basis for explaining cross-lingual differences in word-order learnability, providing a more comprehensive predictor than categorical typological classifications. Vocabulary structure is an integral part of interpreting model surprisal in shuffling experiments.

Future work should examine how model architecture such as tokenization and positional encoding modulate the sensitivity to word-order perturbations, and compare models with human behavior to assess cognitive plausibility. Such research linking language features and model architecture advances the understanding of language learnability.

Limitations

The present study should be interpreted in light of several limitations.

Corpus We use a single high-quality parallel corpus to ensure comparability across languages, yet its number of languages is limited to 21 European languages, of which we selected ten for a focused analysis and computational feasibility. Extending to more corpora would allow for a more diverse set of language typologies to be included (Ploeger et al., 2024, 2025) at the cost of more noise and heterogeneity in the data.

Human learnability We use language models as a tool to study learnability, yet the learnability of a language model does not necessarily generalize to humans. Comparisons with human data, e.g., eye-tracking studies (Schad et al., 2010), could help evaluate cognitive plausibility.

Model size Since our experimental setup requires a large number of models to be trained, the model size is limited in order to achieve reasonable training times. This trade-off could impact vocabulary size effects at very large vocabularies, for which embedding parameters dominate.

Typology We group languages into "free" and "fixed" word order, but typology is a gradient (Levshina et al., 2023; Baylor et al., 2024). A comparative analysis of other continuous typological measures, e.g., subject-object-order entropy, with the vocabulary structure measures we describe here may yield a more nuanced understanding.

Evaluation We evaluate the global surprisal on a test set. An interesting extension would be to assess whether all tokens contribute uniformly or whether surprisal stems can be attributed to breaking certain language-specific collocations, e.g., determiner-adjective-noun constructions.

Ethical considerations

Synthetic languages Our study uses synthetic languages (also called "artificial languages"). There is a wide spectrum of languages, ranging from formal and highly unnatural to attested languages. It is important not to conflate different categories on this spectrum. In our study, we focus on languages that are perturbed only on the dimension of word order, while maintaining the complexity of natural language in terms of lexicon and morphology.

Environmental impact Training models, even with comparatively few parameters, leads to computational cost and CO_2 emissions. We encourage future work to consciously evaluate the need for large-scale studies in order to curtail the ever-increasing environmental impact of our information infrastructure.

Developing and training models for this study used approximately 150 kcore-hours. The models were trained on one A100 with 40 GB with 1400 models for the scan in θ and 600 models for the scan in vocabulary size $|V|$.

Acknowledgments

[Anonymized for review.]

References

Bas Aarts. 2011. *Oxford Modern English Grammar*, 1st edition. Oxford University Press, New York.

Mostafa Abdou, Vinit Ravishankar, Artur Kulmizev, and Anders Søgaard. 2022. Word order does matter and shuffled language models know it. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6907–6919, Dublin, Ireland. Association for Computational Linguistics.

Catherine Arnett and Benjamin K. Bergen. 2025. Why do language models perform worse for morphologically complex languages? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6607–6623, Abu Dhabi, UAE. Association for Computational Linguistics.

Emi Baylor, Esther Ploeger, and Johannes Bjerva. 2024. Multilingual gradient word-order typology from universal dependencies. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 42–49, St. Julian's, Malta. Association for Computational Linguistics.

Lisa Beinborn and Yuval Pinter. 2023. Analyzing cognitive plausibility of subword tokenization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4478–4486, Singapore. Association for Computational Linguistics.

Arianna Bisazza, Ahmet Üstün, and Stephan Sportel. 2021. On the difficulty of translating free-order case-marking languages. In *Transactions of the Association for Computational Linguistics*, volume 9, pages 1233–1248, Cambridge. MIT Press.

Noam Chomsky. 2023. Noam Chomsky: The false promise of ChatGPT. *The New York Times*.

Leshem Choshen and Omri Abend. 2019. Automatically extracting challenge sets for non-local phenomena in neural machine translation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 291–303, Hong Kong, China. Association for Computational Linguistics.

Thomas Hikaru Clark, Clara Meister, Tiago Pimentel, Michael Hahn, Ryan Cotterell, Richard Futrell, and Roger Levy. 2023. A cross-linguistic pressure for uniform information density in word order. In *Transactions of the Association for Computational Linguistics*, volume 11, pages 1048–1065, Cambridge.

Ryan Cotterell, Sebastian J. Mielke, Jason Eisner, and Brian Roark. 2018. Are all languages equally hard to language-model? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 536–541, New Orleans, Louisiana. Association for Computational Linguistics.

Marta Crispino, Cristina Mollica, Valerio Astuti, and Luca Tardella. 2023. Efficient and accurate inference for mixtures of Mallows models with Spearman distance. *Statistics and Computing*, 33(5):98.

William Croft. 2002. *Typology and Universals*, 2nd edition. Cambridge University Press, Cambridge.

Marion Di Marco and Alexander Fraser. 2024. Subword segmentation in LLMs: Looking at inflection and consistency. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12050–12060, Miami, Florida, USA. Association for Computational Linguistics.

Richard Diehl Martinez, David Demitri Africa, Yuval Weiss, Suchir Salhan, Ryan Daniels, and Paula Buttery. 2025. Pico: A modular framework for hypothesis-driven small language model research. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 295–306, Suzhou, China. Association for Computational Linguistics.

Matthew Dryer and Martin Haspelmath. 2024. The World Atlas of Language Structures online - Order of subject, object and verb.

Nadine El-Naggar, Tatsuki Kuribayashi, and Ted Briscoe. 2025a. GCG-based artificial languages for evaluating inductive biases of neural language models. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 540–556, Vienna, Austria. Association for Computational Linguistics.

Nadine El-Naggar, Tatsuki Kuribayashi, and Ted Briscoe. 2025b. Which word orders facilitate length generalization in LMs? An investigation with GCG-based artificial languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35587–35601, Suzhou, China. Association for Computational Linguistics.

William Feller. 1968. *An Introduction to Probability Theory and its Applications I*, 3rd edition. Wiley series in probability and mathematical statistics. Wiley, New York.

Michael A. Fligner and Joseph S. Verducci. 1986. Distance based ranking models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 48(3):359–369.

Víctor Franco-Sánchez, Arnau Martí-Llobet, and Ramon Ferrer-i-Cancho. 2024. Swap distance minimization beyond entropy minimization in word order variation. *Preprint*, arXiv:2404.14192.

Richard Futrell, Edward Gibson, and Roger P. Levy. 2020. Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, 44(3):e12814.

Richard Futrell and Kyle Mahowald. 2025. How Linguistics Learned to Stop Worrying and Love the Language Models. *Behavioral and Brain Sciences*, pages 1–98.

Joseph H. Greenberg. 1990. Some universals of grammar with particular reference to the order of meaningful elements. In Keith Denning and Suzanne Kemmer, editors, *On Language: Selected Writings of Joseph H. Greenberg*, pages 40–70. Stanford University Press.

Ximena Gutierrez-Vasques, Christian Bentz, and Tanja Samardžić. 2023. Languages through the looking glass of BPE compression. *Computational Linguistics*, 49(4):943–1001.

Michael Hahn and Yang Xu. 2022. Crosslinguistic word order variation reflects evolutionary pressures of dependency and information locality. *Proceedings of the National Academy of Sciences*, 119(24):e2122604119.

Coleman Haley and Colin Wilson. 2021. Deep neural networks easily learn unnatural infixation and reduplication patterns. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 427–433, Online. Association for Computational Linguistics.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2025. Glottolog 5.2.

Robert T. Harms. 1997. *Estonian Grammar*, 1st edition. Taylor and Francis, London.

Martin Harris and Nigel Vincent. 2012. *Romance Languages*, 1st edition. Routledge Language Family Series. Taylor and Francis, Hoboken.

Martin Haspelmath. 2023. Defining the word. *WORD*, 69(3):283–297.

John A. Hawkins. 2014. *Cross-Linguistic Variation and Efficiency*, 1st edition. Oxford University Press, Oxford.

Philip Holmes and Ian Hinchliffe. 2013. *Swedish: A Comprehensive Grammar*, 3rd edition. Routledge, London.

Johannes Kabatek. 2022. *Manual of Brazilian Portuguese Linguistics*, 1st edition. Number 21 in Manuals of Romance Linguistics. Walter de Gruyter GmbH, Berlin/Boston.

Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. 2024. Mission: Impossible language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14691–14714, Bangkok, Thailand. Association for Computational Linguistics.

Julie Kallini and Christopher Potts. 2025. Language models as tools for investigating the distinction between possible and impossible natural languages. *Preprint*, arXiv:2512.09394.

Fred Karlsson. 2017. *Finnish: A Comprehensive Grammar*, 1st edition. Routledge, Abingdon, Oxon.

Roni Katzir. 2023. Why large language models are poor theories of human linguistic cognition: A reply to Piantadosi. *Biolinguistics*, 17:e13153.

Maurice G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81.

Istvan Kenesei, Robert M. Vago, and Anna Fenyvesi. 2002. *Hungarian*, 1st edition. Routledge, London.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Alexander Koplenig, Sascha Wolfer, and Peter Meyer. 2023. A large quantitative analysis of written language challenges the idea that all languages are equally complex. *Scientific Reports*, 13(1):15351.

Evelina Leivada, Raquel Montero, Paolo Morosi, Natalia Moskvina, Tamara Serrano, Marcel Aguilar, and Fritz Guenther. 2025. Large language model probabilities cannot distinguish between possible and impossible language. *Preprint*, arXiv:2509.15114.

Natalia Levshina, Savithry Namboodiripad, Marc Allassonnière-Tang, Mathew Kramer, Luigi Talamo, Annemarie Verkerk, Sasha Wilmoth, Gabriela Garrido Rodriguez, Timothy Michael Gupton, Evan Kidd, Zoey Liu, Chiara Naccarato, Rachel Nordlinger, Anastasia Panova, and Natalia Stoynova. 2023. Why we need a gradient approach to word order. *Linguistics*, 61(4):825–883.

Siqi Liu, Jianwei Yan, and Haitao Liu. 2025. The complexity trade-off between morphological richness and word order freedom in Romance languages: A quantitative perspective. *Zeitschrift für romanische Philologie*, 141(2):323–349.

Tom Lundskaer-Nielsen and Philip Holmes. 2015. *Danish: A Comprehensive Grammar*, 2nd edition. Routledge, London.

David J. C. MacKay. 2019. *Information Theory, Inference, and Learning Algorithms*, 22nd edition. Cambridge University Press, Cambridge.

Colin L. Mallows. 1957. Non-null ranking models. I. *Biometrika*, 44(1/2):114–130.

Sebastian J. Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2019. What kind of language is hard to language-model? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4975–4989, Florence, Italy. Association for Computational Linguistics.

James D. Naughton. 2008. *Czech: An Essential Grammar*, 2nd edition. Essential grammars. Routledge, London.

Julie Nijs, Freek Van de Velde, and Hubert Cuyckens. 2025. Is word order responsive to morphology? Disentangling cause and effect in morphosyntactic change in five Western European languages. *Entropy*, 27(1):53.

Steven T. Piantadosi. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21(5):1112–1130.

Steven T. Piantadosi. 2024. Modern language models refute Chomsky's approach to language. In *From Fieldwork to Linguistic Theory*, 1st edition, number 15 in Empirically Oriented Theoretical Morphology and Syntax, pages 353–414. Language Science Press, Berlin.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2012. The communicative function of ambiguity in language. *Cognition*, 122(3):280–291.

Esther Ploeger, Wessel Poelman, Miryam de Lhoneux, and Johannes Bjerva. 2024. What is "typological diversity" in NLP? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5681–5700, Miami, Florida, USA. Association for Computational Linguistics.

Esther Ploeger, Wessel Poelman, Andreas Holck Høeg-Petersen, Anders Schlichtkrull, Miryam De Lhoneux, and Johannes Bjerva. 2025. A principled framework for evaluating on typologically diverse languages. *Computational Linguistics*, pages 1–36.

Wessel Poelman, Thomas Bauwens, and Miryam de Lhoneux. 2025. Confounding factors in relating model performance to morphology. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7273–7298, Suzhou, China. Association for Computational Linguistics.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Dace Praulinš. 2012. *Latvian: An Essential Grammar*, 1st edition. Essential grammars. Routledge, London.

Geoffrey Sampson. 2009. A linguistic axiom challenged. In Geoffrey Sampson, David Gil, and Peter Trudgill, editors, *Language Complexity as an Evolving Variable*, pages 1–18. Oxford University PressOxford.

Geoffrey Sampson, David Gil, and Peter Trudgill, editors. 2009. *Language Complexity as an Evolving Variable*. Oxford University PressOxford, New York.

Daniel J. Schad, Antje Nuthmann, and Ralf Engbert. 2010. Eye movements during reading of randomly shuffled text. *Vision Research*, 50(23):2600–2616.

Thomas Schürmann and Peter Grassberger. 1996. Entropy estimation of symbol sequences. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 6(3):414–427.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Anna Siewierska, editor. 2010. *Constituent Order in the Languages of Europe*, 1st edition. Number Eurotyp 20-1 in Empirical Approaches to Language Typology. De Gruyter, Berlin.

Taiga Someya, Anej Svete, Brian DuSell, Timothy J. O'Donnell, Mario Giulianelli, and Ryan Cotterell. 2025. Information locality as an inductive bias for neural language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27995–28013, Vienna, Austria. Association for Computational Linguistics.

Helena Sulkala and Merja Karjalainen. 2012. *Finnish*, 1st edition. Descriptive grammars. Routledge, London.

Peter Svenonius. 2025. Word order universals and their relationship to structure. *Annual Review of Linguistics*, 11(1):137–162.

Wenpin Tang. 2019. Mallows ranking models: Maximum likelihood estimate and regeneration. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6125–6134. Proceedings of Machine Learning Research.

Chaofan Tao, Qian Liu, Longxu Dou, Niklas Muennighoff, Zhongwei Wan, Ping Luo, Min Lin, and Ngai Wong. 2024. Scaling laws with vocabulary: Larger models deserve larger vocabularies. In *Advances in Neural Information Processing System*, volume 37, pages 114147–114179. Curran Associates, Inc.

Tianyang Xu, Tatsuki Kuribayashi, Yohei Oseki, Ryan Cotterell, and Alex Warstadt. 2025. Can language models learn typologically implausible languages? *Preprint*, arXiv:2502.12317.

Xiulin Yang, Tatsuya Aoyama, Yuekun Yao, and Ethan Wilcox. 2025. Anything goes? A crosslinguistic study of (im)possible language learning in LMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26058–26077, Vienna, Austria. Association for Computational Linguistics.

George Kingsley Zipf. 1935. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*, 1st edition. Houghton Mifflin, Boston, Massachusetts.

George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley Press., Cambridge.

Imry Ziv, Nur Lan, Emmanuel Chemla, and Roni Katzir. 2025. Biasless language models learn unnaturally: How LLMs fail to distinguish the possible from the impossible. *Preprint*, arXiv:2510.07178.

A Shuffling algorithm

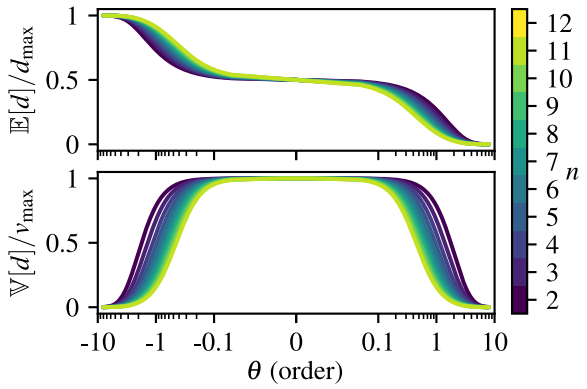

Figure 7: Normalized analytical mean and variance of Kendall's τ with Mallows shuffling over the order θ for different sentence lengths n.

The idea of sampling one permutation π for each sentence length n to shuffle a language corpus deterministically has been used in (Someya et al., 2025). We introduce the Mallows model (Tang, 2019) as a unifying probabilistic method for selecting the permutation π for each sentence length n.

The Mallows model assigns the probability of a permutation $\pi \in \mathfrak{S}_n$, based on the original word

Table 1: Language selection with name and ISO-code abbreviation, grouped by branch and family (IE: Indo-European). We list word-order flexibility and morphology.

Language	ISO	Branch	Family	Flexibility	Morphology	Reference
English	en	Germanic	IE	fixed	analytic	(Aarts, 2011)
Danish	da				analytic	(Lundskaer-Nielsen and Holmes, 2015)
Swedish	sv				analytic	(Holmes and Hinchliffe, 2013)
Portuguese	pt	Romance			fusional	(Kabatek, 2022; Harris and Vincent, 2012)
French	fr				fusional	(Harris and Vincent, 2012)
Latvian	lv	Baltic	IE	free	fusional	(Praulinš, 2012)
Czech	cs	Slavic	IE		fusional	(Naughton, 2008)
Hungarian	hu	Ugric	Uralic		agglutinative	(Kenesei et al., 2002)
Estonian	et	Finnic			agglutinative	(Harms, 1997)
Finnish	fi				agglutinative	(Sulkala and Karjalainen, 2012; Karlsson, 2017)

order $\pi_0 = (1, 2, \ldots, n)$, as

$$\mathbb{P}_{\theta,\pi_0,d}(\pi) = \frac{1}{Z(\theta, d)} \mathrm{e}^{-\theta d(\pi,\pi_0)} \tag{3}$$

where the control parameter θ is the order (also called dispersion or concentration (Crispino et al., 2023), analogous to an inverse temperature β), d is a distance metric measuring the discrepancy between π and π_0, and Z is the partition function that normalizes the distribution. The order θ is interpreted as how preferred the original order π_0 is by the probability distribution.

Since the Mallow's ϕ model (Tang, 2019) is easy to sample from (see (Fligner and Verducci, 1986) for details)l, we use Kendall's τ as distance metric (Kendall, 1938),

$$d_\tau(\pi \circ \pi_0^{-1}) = \mathrm{inv}(\pi \circ \pi_0^{-1}) \tag{4}$$

where $\mathrm{inv}(\pi) := |\{(i, j) \in [n]^2 : i < j \wedge \pi(i) > \pi(j)\}|$, that is, d_τ is the minimum number of adjacent swaps to restore the central order π_0 from the permutation π.

According to (Fligner and Verducci, 1986), the Mallow's tau model has the mean (for permutations $\pi \in \mathfrak{S}_n$ of length n)

$$\mathbb{E}_\theta(d_\tau) = \frac{n\mathrm{e}^{-\theta}}{1 - \mathrm{e}^{-\theta}} - \sum_{j=1}^{n} \frac{j\mathrm{e}^{-j\theta}}{1 - \mathrm{e}^{-j\theta}} \tag{5}$$

and variance

$$\mathbb{V}_\theta(d_\tau) = \frac{n\mathrm{e}^{-\theta}}{(1 - \mathrm{e}^{-\theta})^2} - \sum_{j=1}^{n} \frac{j^2\mathrm{e}^{-j\theta}}{(1 - \mathrm{e}^{-j\theta})^2} \tag{6}$$

with maximum distance (Kendall, 1938)

$$d_{\max} = \binom{n}{2} = \frac{n(n-1)}{2} \tag{7}$$

between permutations and maximum variance (Feller, 1968, p. 257)

$$v_{\max} = \frac{n(n-1)(2n+5)}{72}, \tag{8}$$

respectively.

Figure 7 shows the normalized mean and variance of the Mallows ϕ distribution by the order θ for different sentence lengths n.

B Language selection

We select ten out of the 21 languages available in Europarl for our experiments: Five languages commonly classified—either categorically or via continuous measures (Siewierska, 2010; Levshina et al., 2023)—as fixed-word-order and five as free-word-order, namely: English, Danish, Swedish (Germanic), French, Portuguese (Romance), Latvian (Baltic), Czech (Slavic), Hungarian, Estonian, Finnish (Finno-Ugric), see Table 1. The dominant or neutral word order of all these languages is SVO (Siewierska, 2010).

C Preprocessing

We cleaned the raw Europarl sentences prior to shuffling using these steps.

- Remove empty or punctuation-only sentences, speaker and language labels, and obvious non-speech content.
- Fix or remove Unicode artifacts: replace soft hyphens (U+00AD) with "-"; remove replacement characters (U+FFFD) and zero-width spaces (U+200B); drop lines containing URLs.
- Strip nested parenthetical and bracketed content, quotation marks, and apostrophes while preserving enclosed text.

- Normalize punctuation: remove stray commas, split at semicolons and colons outside words; clean bullets and dashes, replacing with hyphens where appropriate.
- Collapse whitespace, lowercase text, and ensure terminal punctuation.
- Apply minimal language-specific rules: remove mistaken spaces in Finnish abbreviations ("EU: n" → "EU:n").

For each language, we remove sentences longer than 80 words, and then split into training, validation, and test sets of 650 000, 5000, and 5000 sentences, respectively.

Full preprocessing code and regex rules are available at [GitLab link anonymized for review].

D Training parameters

Table 2: Hyperparameters for the PICOLM models used in our experiments.

Parameter	Value
Architecture	Transformer decoder
Total parameters	50.5 M
Layers	12
Embedding size	384
Attention heads	12
Attention KV heads	4
Hidden dimension	1536
Sequence length	512
Tokenizer	ByteLevel BPE
Tokenizer min. freq.	2
Vocabulary size $\lvert V\rvert$	16 000 / varied
Optimizer	AdamW
Learning rate	0.0014
Learning rate schedule	Linear
Warmup steps	5
Batch size (training)	64
Training steps	1000
Order θ	varied / $\{0, 9\}$

Table 2 lists the hyperparameters of training the language models for our experiments. Each model was trained on one A100 GPU.

We generate five random seeds per language variant and apply the deterministic word-level shuffling for a range of orders θ, training each model with batch size 64 for 1000 steps. The vocabulary size is $|V| = 16\,000$ when varying the order θ, with roughly log-scaled $\theta \in [-9, 9]$. When the vocabulary size is varied as $|V| = \{258, 1000, 8000, 16\,000, 32\,000, 64\,000\}$ the order is chosen as $\theta \in \{0, 9\}$. Note that $256 + 2$ corresponds to character-level tokenization.

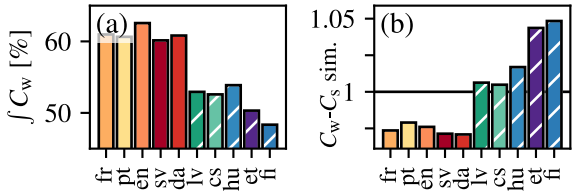

Figure 8: Coverage measures: The (a) word coverage integral and (b) word-subword-coverage similarity.

E PLS regression

Here, we define and list predictors used for the partial-least-squares (PLS) regression analysis.

E.1 Definition of coverage metrics

We calculate the coverage integral as the area under the word coverage curve C_w per log-rank up to $r_{max} = 10^5$:

$$\frac{1}{\log(r_{max})} \int_1^{r_{max}} C_w \log(\mathrm{d}r)\,. \qquad (9)$$

The coverage similarity relates word and subword coverage through a regression slope m in log-space without intercept,

$$m = \frac{\sum_{r=1}^{r_{max}} w_r\, C_w(r)\, C_s(r)}{\sum_{r=1}^{r_{max}} w_r\, (C_w(r))^2} \qquad (10)$$

with weights given by $\log(r)$.

Both coverage integral and similarity, visualized in Fig. 8, clearly separate free- from fixed-word-order languages.

E.2 Regression factors

All vocabulary statistics that we use as predictors for the PLS analysis are listed by language in Table 3 along with the classification into free- and fixed-word-order. The predictors are grouped by coverage, sentence length, and morphological complexity.

The morphological complexity metrics are: fertility, i.e., the average number of subwords per word; word length in characters; and types in the sense of unique words, i.e., the word vocabulary of the corpus.

E.3 Correlation of factors

The correlation matrix in Fig. 9 shows that all predictors are highly correlated, motivating the use of partial-least-squares regression.

Table 3: Vocabulary statistics for each language: Word and subword coverage at rank 100, coverage similarity and integral; sentence length in subwords and words; morphological complexity measured by fertility, word length in characters, and word types.

		Coverage				Sentence length		Morph. complexity		
Lang.	Order	$C_{\mathrm{w},100}$	$C_{\mathrm{s},100}$	C simil.	$\int C_{\mathrm{w}}$	Subw./sent.	Words/sent.	Fertility	Word len.	Types
fr		52.7	49.0	0.974	61.0	28.1	24.4	1.15	6.02	96 727
pt		51.0	47.9	0.979	60.6	28.1	24.2	1.16	6.03	108 442
en	Fixed	55.4	52.6	0.976	62.6	26.4	23.8	1.11	5.70	70 536
sv		52.6	47.5	0.971	60.1	24.4	20.4	1.20	6.22	177 002
da		54.2	49.6	0.971	60.8	25.7	21.7	1.19	6.09	179 915
lv		38.3	35.7	1.006	52.9	23.1	18.2	1.27	6.88	156 845
cs		37.1	34.0	1.005	52.6	24.8	19.6	1.27	6.35	169 003
hu	Free	40.9	36.6	1.017	53.9	24.8	18.7	1.33	7.23	307 197
et		35.5	33.0	1.044	50.3	22.1	16.5	1.34	7.40	283 165
fi		33.5	29.8	1.048	48.3	23.2	16.2	1.43	8.33	363 154

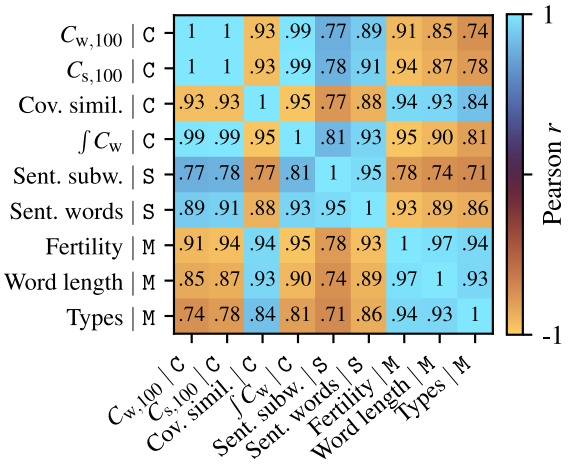

Figure 9: Correlation matrix of the predictors, grouped by coverage (C), sentence length (S), and proxies of morphological complexity (M).

E.4 Latent components

Figure 10 (a) shows the two components identified by the PLS regression: The vocabulary component loads on all predictors, but more strongly on coverage; the morphological-complexity component loads primarily on fertility, word length, and word types.

In panel (b), we see that the vocabulary component is structurally uniform across the spectrum of θ, whereas the morphological-complexity component aligns most with the irregular word order at small $|\theta|$.

Predicting per θ only with the single binary predictor of free vs. fixed word order yields $R^2 = 0.44$ to 0.85 with mean $\bar{R}^2 = 0.65$ and lowest per language $R^2 = 0.77$ for Portuguese and overall $R^2 = 0.94$. A non-predictive feature like characters per token yields per-θ mean $\bar{R}^2 = -0.02$ and overall $R^2 = 0.85$. The R^2 of the latter remains relatively high because of the shared structure in the $S(\theta)$ curves. We observe that word-order flexibility is not as comprehensive as the combination of vocabulary structure metrics, but explains more than the weak predictor of characters per token.

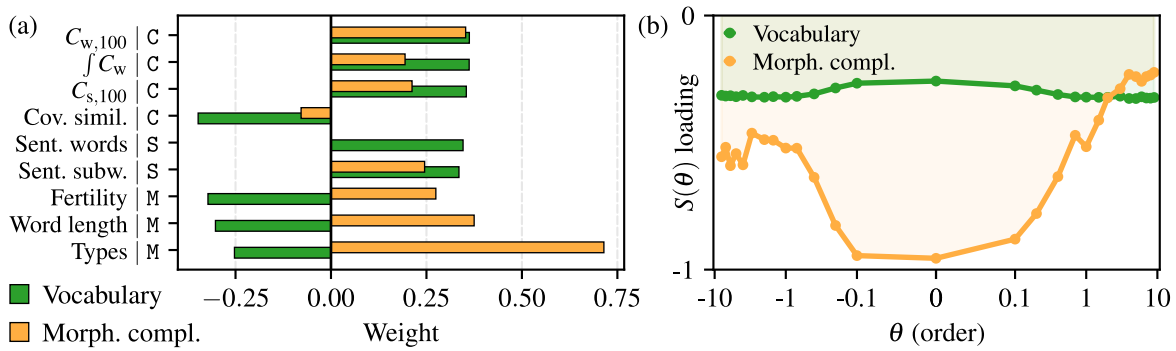

Figure 10: (a) Predictor and (b) response loadings of the vocabulary and morphological complexity component.