Do LLMs learn a true syntactic universal?

John T. Hale
Google DeepMind
Johns Hopkins University
jthale@google.com

Miloš Stanojević
Google DeepMind
University College London
stanojevic@google.com

Abstract

Do large multilingual language models learn language universals? We consider a much discussed candidate universal, the Final-over-Final Condition (Sheehan et al., 2017b). This Condition is syntactic in the sense that it can only be stated by reference to abstract sentence properties such as nested phrases and head direction. A study of typologically diverse "mixed head direction" languages confirms that the Condition holds in corpora. But in a targeted syntactic evaluation, Gemini Pro only seems to respect the Condition in German, Russian, Hungarian and Serbian. These relatively high-resource languages contrast with Basque, where Gemini Pro does not seem to have learned the Condition at all. This result suggests that modern language models may need additional sources of bias in order to become truly human-like, within a developmentally-realistic budget of training data.

1 Introduction

The question of whether large language models (LLMs) display human-level competence has provoked lively discussion. Some commentators' minds are made up: "LLMs have already demonstrated that human-like grammatical language can be acquired without the need for a built-in grammar" (Contreras Kallens et al., 2023). Others are more guarded: "nearly all studies have reported that [deep neural networks'] behavior deviated from the idealized syntactic competence that a linguist might postulate" (Linzen and Baroni, 2021).

This controversial question is important both for continued progress in AI systems, as well as for debates over inborn biases that might be necessary to achieve human-level language within a developmentally-plausible number of training examples (Warstadt et al., 2023). Typological generalizations are a key battleground in such debates (see e.g. van der Hulst, 2023, chapter 7). Where such generalizations qualify as language universals, they could potentially underwrite a strong nativist argument for inborn biases (e.g. Chomsky, 1965, 25). But this could only happen if competing explanations, based upon general-purpose learning rules, were ruled out. Can LLMs rebut nativism by accounting for true language universals without universal grammar? That is the question this paper takes up.

Our answer is negative. We find that modern LLMs *do* learn a candidate syntactic universal if given superhuman amounts of text. But in the Basque language, where data sizes are likely closer to human level, these same models do not learn the universal. This outcome leaves nativism still standing, because the attained level of performance does not meet the criteria of universality or developmental plausibility.

Section 2 begins by introducing a candidate language universal in the domain of syntactic structure called the Final-over-Final Condition (Holmberg, 2000; Biberauer et al., 2014; Sheehan et al., 2017b). Section 3 reports a corpus study of this universal in six "mixed head direction" languages where it conceivably could be violated. The corpus study confirms that the universal indeed holds in these languages, one of which is Basque. Section 4 goes on to a targeted syntactic evaluation of PaLM (Chowdhery et al., 2023) and Gemini Pro (Gemini Team, 2023). Gemini is, at the time of writing of this paper, the most advanced of Google's large language models. It surpassed GPT-4 on the LMSYS.org leaderboard on January 26th 2024. PaLM is an earlier model. Both learned the Condition in all languages except Basque. Section 5 discusses this gap between human and model performance, noting that neither data size nor parameter count matter, beyond some threshold point.

We contribute the first large-scale corpus study of this universal as well as a novel evaluation that speaks to human language in general.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 17106–17119

2 The Final-over-Final Condition

The Final-over-Final Condition (henceforth: FOFC) is "either a very strong tendency among the languages of the world or actually a language universal" (Sheehan et al., 2017b, page 2). Here we do not take FOFC as a given fact, but as a hypothesis in need of testing.

FOFC depends on the notion of head direction. Larson's 2010 textbook introduces this notion with two examples, reproduced below from page 351.

(1) a. Homer may leave.
b. Homer may visit Marge.
c. Homer may give an apple to Marge.

(2) a. Taroo-wa deru daroo.
T.-TOP depart may
b. Taroo-wa Hanako-o tazuneru daroo.
T.-TOP H.-ACC visit may
c. Taroo-wa ringo-o Hanako-ni ageru daroo.
T.-TOP apple-ACC H.-DAT give may

English, as exemplified in (1), contrasts with Japanese in example (2) along the dimension of head direction. The three sub-examples for each language show that there exists a variable-length unit – the verb phrase (VP) – that may contain zero, one or two non-head phrases called (syntactic) complements. In English the verb precedes its complements, whereas in Japanese it follows them. This situation is summarized by saying English has head-initial VPs, whereas in Japanese VPs are head-final. Various criteria have been proposed to differentiate heads from non-heads in a phrase. Heads typically determine the part of speech or morphological form of their dependent (see Bender 2013, number 52 or Zwicky 1985). On the basis of these examples, Larson goes on to remark upon another grammatical category, T. His observation rests on a common categorization of the English modal "may" and the Japanese future-possibility word "daroo" as T, a category that includes tense and other auxiliary verbs. Analogous to VP, tense phrases TP are head-initial in English but head-final in Japanese. This basic idea of viewing phrasal head direction as a language-specific parameter setting is introduced on pages 73–75 of Stowell (1981). Newmeyer (2005, 43-44) and Sheehan (2021, §11.3) survey subsequent developments of this basic idea.

The FOFC, then, is a very general condition on headed phrase structure trees of depth two. At this depth there are four possible configurations, shown in Figure 1. Two of these configurations, 1a and 1b are "harmonic" in the sense that both superphrase and subphrase follow the same head direction. These structures fit well the facts of English and Japanese, respectively. The other configurations are "disharmonic" such that the head direction of the superphrase is different from that of the subphrase. These phrase structures are useful for analyzing languages like Finnish, where a VP may be head-final in the context of a focused complementizer (Holmberg, 2000, 2017). Although there may be a tendency across time for languages to change in the direction of more consistent, harmonic word orders (see e.g. Gulordava and Merlo, 2015), mixed word order is by no means rare.

The FOFC claim is that one of the two possible disharmonic configurations is universally banned in human language. In Figure 1 the outlawed structure is the grayed-out cell, 1d. This strong claim[1] is tempered somewhat by the requirement that β be an extended projection of α in the sense of Grimshaw (2005, chapter 1). The basic idea, which has enjoyed wide currency since the early 1990s, is to group phrases that are headed by function words together with their main content word, typically a noun or verb. This restriction to extended projections tightens the FOFC so that it applies solely within a given domain i.e. nominal or verbal. Biberauer (2017, 245) grapples further with the sorts of derived structures to which the FOFC applies.

Perhaps the strongest evidence for the FOFC comes from the case where α = V and β = T. Here, the FOFC outlaws a head-final TP over a head-initial VP. As Sheehan et al. (2017a) discuss, evidence from Germanic languages seems to bear this out. These languages attest mixed headedness; in fact virtually every possible combination of auxiliaries, verbs and their objects is attested. What is not attested in the literature on comparative Germanic is a tensed auxiliary verb following a verb-object combination. This would be analyzed as a head-final TP with a head-initial VP daughter, exactly what the FOFC rules out. Section 3 considers this case in quantitative detail.

[1]The FOFC is especially relevant to debates over nativism because of the way it has withstood attempts at functional explanation. Sheehan et al. (2017b, chapter 5) refutes two prominent proposals in this category. Hawkins (2014, chapter 5) takes issue with the empirical claim. Section 3 therefore examines FOFC as an empirical claim.

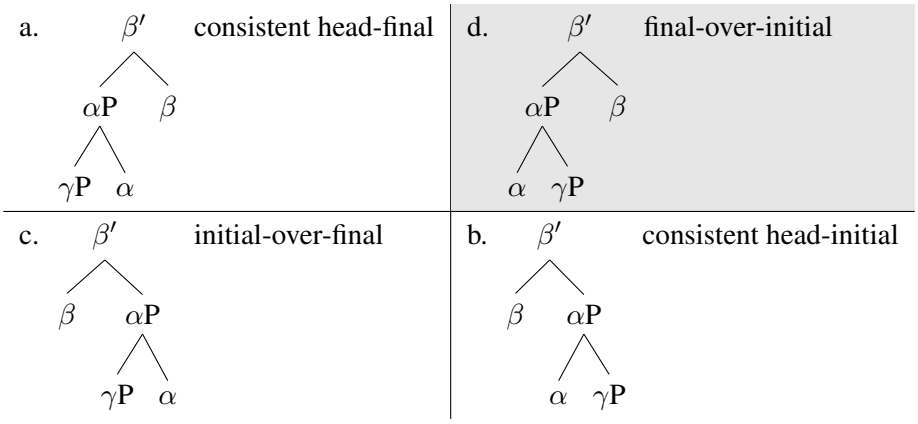


Figure 1: The Final-Over-Final Condition bans head-final superphrases from having head-initial subphrases. α and β are variables that range over grammatical categories such as Noun, Verb, Tense, Determiner etc. These are heads of the relevant superphrase and subphrase, respectively. γP is the syntactic complement of the subphrase; here γ is also a variable over categories. The primed node label β' indicates a phrase that projects from β but need not be a maximal projection, in the sense of X-bar theory. The order of the lettered examples follows Sheehan et al. (2017b).

3 Corpus Study

The corpus study asks whether the FOFC is true or not; Section 4 follows up by asking whether a modern neural language model can distinguish FOFC-respecting from FOFC-flouting sentences. This first question is addressed by asking: does a contingency table for nested combinations of Aux and V look the way the FOFC predicts it should look? Referring to Figure 1, one asks: is the number in cell 1d lower than expected? The corpus study thus seeks "negative" evidence for the grayed out configuration (Stefanowitsch, 2006). This methodology closely follows Merlo (2016) and Gulordava and Merlo (2020) in applying the χ^2 test, a standard procedure in corpus analysis (for a introduction, see Brezina 2018, §4.3 or Stefanowitsch 2020, §6.3). The present study is more detailed than previous FOFC surveys, such as Bazalgette (2012, cited on page 91 of Sheehan et al. 2017b) insofar as it is based upon counts of individual sentences rather than descriptive reports about entire languages.

As section 2 emphasized, the key claim of the FOFC has to do with disharmonic word orders. It does not rule out any structures in consistently head-initial languages like English and Indonesian, or consistently head-final languages like Japanese and Turkish. For this reason, the corpus study focuses on languages that have been described (however superficially) as displaying mixed-headedness. The central comparison is between the permitted initial-over-final order (1c) and the outlawed final-over-initial order (1d). Counts in the other cells are only needed to establish the expected value in cell 1d.

We consider languages collected in the C4/multilingual dataset (Xue et al., 2021), except for German. In German for idiosyncratic reasons we analyzed a dump from wikipedia. The dump happened near the end of 2020 and was processed using wikiextractor (Attardi, 2015). All languages were analyzed up through the stage of dependency parsing with the Stanza toolkit (Qi et al., 2020). In Serbian, we applied a version of Stanza that is optimized for South Slavic languages (Terčon and Ljubešić, 2023). To make parsing more manageable, we excluded outputs from the Stanza sentence-breaker that exceeded 40 words in length.

There is debate among dependency grammarians over the status of function words. The well-known Universal Dependencies (UD) scheme views function words as dependents of content words, on the assumption that content words will prove more helpful in semantic analysis (de Marneffe et al., 2021). This decision conflicts with the view of TP as an extended projection of V – a background assumption in the argument for the FOFC from comparative Germanic. To reverse it, we convert the UD dependency graphs obtained from freely-available Stanza models into Surface UD graphs (Gerdes et al., 2018). Surface UD (SUD) takes the opposite view — in SUD, function words such as auxiliary verbs are heads (not dependents). Their dependents are the same lexical verbs that would have been their heads in UD. It is straightforward

language	sentence count	χ^2	significance level
Hungarian	182M	6498	$p < 10^{-16}$
Basque	41M	1499197	$p < 10^{-16}$
Russian	483M	290076	$p < 10^{-16}$
Serbian	124M	14906	$p < 10^{-16}$
German	20M	too few attestations for chi-squared test	

Table 1: Corpus study results. All languages (save for German) show a statistically-significant effect of the Final-over-Final Condition. Each row reports the value of a χ^2 statistic, quantifying how far away a contingency table such as Table 2 is from the values one would expect if Verb-Object and Aux-VP order were independent. See main text regarding German.

to view the converted SUD dependency graphs as X-bar trees, as envisaged in the FOFC. The UD/SUD part of speech tag Aux corresponds best to category T as used in generative grammar.

The results in each case take the form of a contingency table. For the sake of space, we discuss only Hungarian, Basque, German and Serbian, summarizing results from the other languages in Table 1. The contingency table for Hungarian is given in Table 2. In this table, the less-than sign denotes linear precedence. The columns of the table show linear order possibilities within the verb phrase (VP). In these column headers 'O' denotes the object of a verb – its syntactic complement, regardless of category. The first column corresponds to head-finality, the second to head-initiality. The rows identify linear orders for the Auxiliary with respect to its complement VP.

O<V	V < O	
4401	320	VP < Aux
9530	20754	Aux < VP

Table 2: Hungarian two-phrase configurations {Aux, V} in sentences of length 40 or less.

The counts in Table 2 are obtained by searching SUD dependency graphs using the sort of query shown in Appendix B. The pattern suggests that the head-direction of VP and AuxP are not independent in Hungarian. The FOFC-violating configuration is attested just 320 times which is far fewer than the expected value, 2842.176. Indeed it is the rarest of all four configurations. Manual examination, elaborated below, suggests in every case that we have examined, that this residue is attributable to analysis errors.

Parser error inevitably plays a role in a large-scale study such as this. In an effort to reduce parser error we performed a small study of Hungarian sentences of 12 words or less. This yielded the same pattern, with 34 examples in the FOFC-violating cell. Professor Tibor Laczkó, an expert on Hungarian syntax who is also a native speaker, examined these examples and determined that none of them are true FOFC violations. The most well-attested error types were:

1. topicalization: the V or the O has moved to a sentence-medial position where it is no longer part of the extended projection of Aux
2. mistagging a homograph (e.g. the word that means "tooth") as an auxiliary verb. This same form occurs certain fixed expressions that do not form AuxPs.
3. mis-attaching a verb in a preceding if-clause to a lower auxiliary verb
4. focus: the V or the O has moved to a sentence-medial position where it is no longer part of the extended projection of Aux
5. sentence segmentation errors

Error types 1 and 4 involve movements which break the relationship of extended projection between T and V. This renders the FOFC inapplicable, as suggested earlier on page 2. A query for finding these Hungarian cases is given in Appendix B.

We applied Stanza to all the Basque text in C4/multilingual; these results are shown in Table 3. Professor Ricardo Etxepare, an expert on Basque syntax who is also a native speaker, manually examined a sample of 28 cases out of the 1632 putative FOFC violations. None actually violated the FOFC. The most well-attested error types were:

O<V	V < O	
7099566	1632	VP < Aux
291281	79119	Aux < VP

Table 3: Basque two-phrase configurations {Aux, V} in sentences of 40 words or less.

1. unacceptable string (9)
2. attachment error (7, often with relativization)
3. tagger error (4)
4. sentence segmentation error (2)

In German, we restrict consideration to embedded clauses. Word order in embedded clauses is not disrupted by movement to second position, as it is in main clauses. To find such cases, we add the requirement that the auxiliary verb be governed by a complementizer "dass" via an arc labeled `comp:obj`. The results are shown below in Table 4.

O<V	V < O	
34498	13	VP < Aux
74	3	Aux < VP

Table 4: German two-phrase configurations of {Aux, V} in embedded clause.

The paucity of verb-initial examples in embedded clause renders the χ^2 test inapplicable. This paucity is consistent with the usual characterization of German as underlyingly verb-final (for a textbook treatment see e.g. Müller, 2023). Professor Vera Lee-Schoenfeld, an expert on German syntax and a native speaker, manually examined the thirteen examples in the FOFC-outlawed configuration and determined that none were true FOFC violations. The three main types of error were:

1. mistagging a prenominal adjective as V
2. mistagging a participle at the end of a reduced relative clause, or a finite verb at the end of a full relative clause as V
3. mistagging a noun or part of a compound noun as V

One of the more notable verb-initial examples came from a sacred song that is cited in Wikipedia; the song itself is dated 1529 but it may have been written earlier. This underlines the point that modern German does not attest head-initial VPs in embedded clause. For this reason, and given the lack of controversy surrounding the FOFC in German (see last paragraph of section 2) we did not proceed with further parsing or analysis of this language.

Two additional constraints were applied to the Slavic languages, Russian and Serbian, in order to rule out VP fronting as detailed in Appendix C. In Serbian the same pattern manifests itself, shown below in Table 5. One of us examined a sample of 30 cases out of the 2197 putative FOFC violations. Again, none were actual violations. The most frequent error types were:

O<V	V < O	
2442	2197	VP < Aux
13928	212251	Aux < VP

Table 5: Serbian two-phrase configurations {Aux, V} in sentences of 40 words of less.

1. attachment error
2. sentence segmentation

The FOFC could have been disconfirmed by finding roughly equal numbers of structures like 1c and 1d, but that was not what we observed. There are far fewer instances of the configuration that the FOFC bans, 1d, than would be expected by chance alone. When small samples of these spurious matches are examined by experts, none of them turn out to be actual violations. From this we tentatively conclude that the FOFC holds, and proceed to ask whether modern neural language models can learn this constraint.

4 Evaluating LLMs on the FOFC

The targeted syntactic analysis reported in this section asks whether a language model can what a human can do. What humans can do is illustrated by the combination of German judgments in 3 on page 6. This pattern, cited by Biberauer (2017, 245), shows the influence of the FOFC. Consider first the preliminary example (3a). This example shows that Verb-Object order is possible in Colloquial German; it is possible in a VP-topicalized main clause context. In this kind of derived structure, the Aux is not an extended projection of VP. For this reason, we sought to exclude such structures in the corpus studies of German, Serbian and Russian. However in embedded clauses where verb second movement does not apply, Subject-Verb-Object-Aux order leads to unacceptability as shown in example (3b). This is the FOFC effect. The example can be saved (3c) by extraposing the Object "mit ihr." The fact that (3b) is unacceptable whereas (3c) is accepted by competent German speakers may be understood as reflecting the influence of the FOFC.

The targeted syntactic analysis evaluates whether Gemini Pro and PaLM show an influence of the FOFC along the lines of the human judgments in Figure 2 by scoring minimal pairs (see e.g. Warstadt et al., 2019; Niu and Penn, 2020; Marvin

(3) a. [Gesprochen [mit ihr]]$_{VP}$ hat$_{Aux}$ er nicht mehr
spoke to her has he not more
'As for speaking with her, he no longer did that'

b. ∗ dass er nicht mehr [gesprochen mit ihr] hat
that he not more spoken with her has

c. dass er nicht mehr [gesprochen hat] mit ihr
that he not more spoken has with her

Figure 2: Effect of the FOFC on German acceptability (Haider, 2012, 80). S V O Aux word order in (3b) leads to unacceptability (marked with asterisk) and must be repaired, for instance by extraposing the Object outside of VP to the end of the sentence as shown in (3c).

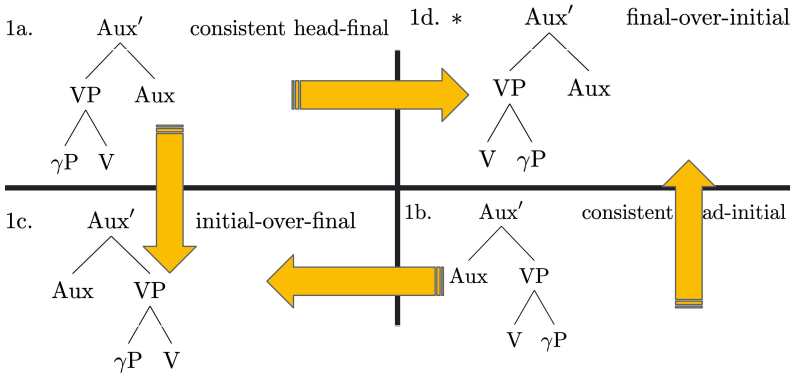


Figure 3: Synthesizing minimal pairs by transformation. See Appendix A for code.

and Linzen, 2018; Sprouse et al., 2018; Lau et al., 2017). These minimal pairs are created by transforming attested examples from C4/multilingual as shown in Figure 3.

The transformations reorder dependency subtrees so as to create strings that exemplify disharmonic word orders. Such reordering preserves the length, lexical content and (parser-inferred) grammatical relations of the example. Indeed its impact on acceptability need not be disastrous. For instance, changing from verb-initial to verb-final order (i.e. 1a to 1c in the sense of Figures 1 and 3) transforms an example from the English Web Treebank, "Google finally had an analyst day" into "Google finally an analyst day had." In the estimation of one author, this transformed sentence sounds a bit archaic but is not total word salad. We measure the level of respect for the FOFC by subtracting the log-probability of 1c-type examples (FOFC-respecting) minus the log-probability of 1d-type (FOFC-disrespecting) examples.

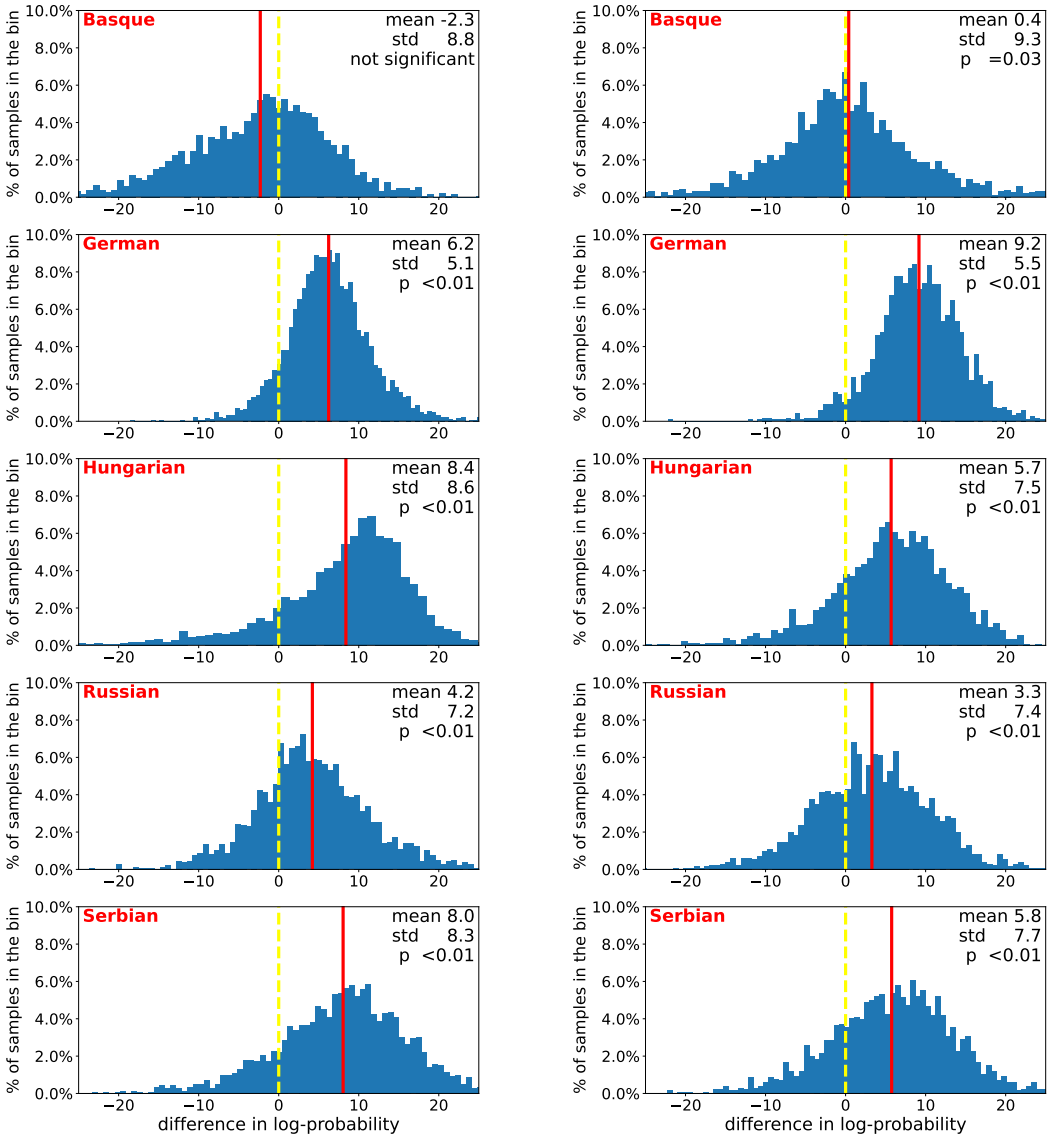
The variability in bell-shaped distributions shown in this Figure 4 could reflect errors parsing the untransformed example, information structure factors like focus, or register variation of an essentially sociolinguistic nature. By examining a large sample of matched examples, the analysis seeks to average out this noise and measure the degree of respect the FOFC, per se.

5 Results and Discussion

The average penalty values in Figure 4 can be interpreted as acceptability differences. In Hungarian, Russian, Serbian and German the distribution is centered above zero for both models. This suggests that Gemini Pro and PaLM have learned the FOFC in those languages. However Basque is clearly different. In Basque, Gemini Pro assigns rather more probability to FOFC-violating sentences than to FOFC-compliant ones. PaLM is only slightly better in terms of mean value for Basque, but most of the distribution's penalties are still in the negative region.

The corpus study from section 3 suggest that the FOFC does hold in Basque, at least in C4/multilingual. These findings imply that Gemini Pro and PaLM have not reliably learned the Final-Over-Final Condition in Basque.

There are at least two possible explanations for this phenomenon: model size (i.e. parameter count) and data size (i.e. example count). At this time,



(a) Gemini Pro.

(b) PaLM 8B.

Figure 4: Results for Gemini Pro and Palm 8B. Each graph shows the distribution of FOFC violation penalties for minimal pairs as described in Section 4. The x-axis is the difference in log-probabilities of FOFC abiding initial-over-final order and FOFC breaking final-over-initial order. The y-axis is, as usual with histograms, the ratio of samples that fall in a range represented by a bar. The red line marks the empirical distribution's mean value, which will be to the right of 0 (dashed yellow line) if the model prefers FOFC-respecting over FOFC-flouting examples. P-values are from the single-tailed t-test for hypothesis that the mean is > 0. The pattern for both LLMs look similar: in German, Hungarian, Russian and Serbian they learn FOFC constraint, while Basque proves challenging. Mean value is slightly better for PaLM 8B, but still the majority of distribution falls into the negative region.

Google has not disclosed either of these sizes for Gemini Pro, but for PaLM model all the details of training data, model size and evaluation are provided in the technical report by Chowdhery et al. (2023). We have tested all PaLM sizes up to half-trillion parameters version. Due to space constraints, results for other sizes of PaLM are presented in in Appendix in Figure 5. The pattern is largely the same as with Gemini Pro and PaLM 8B.

5.1 A training data threshold

Work by Davis (2022) and others emphasizes the limitations of training data in explaining observed mismatches between language models and human speakers. This perspective invites scrutiny of the data sizes for the languages at issue; these are shown for PaLM in Table 6. There is no clear correlation between these training data sizes and the magnitude of the model-assigned penalties. How-

language	tokens
Basque	153M
German	25.954B
Hungarian	555M
Russian	3.932B
Serbian	373M
Croatian	198M
Bosnian	427M

Table 6: Size of PaLM training data for a subset of languages as measured in the number of tokens taken from Chowdhery et al. (2023). Croatian and Bosnian counts are provided for comparison with Serbian, as discussed in the main text.

ever, it is clear that the only language for which FOFC is challenging to learn is the one with the least training data. That language, Basque, is the the closest to the threshold of developmental plausibility cited by Warstadt et al. (2023), 100M tokens. The next least-resourced language, that we tested, Serbian, might seem like a borderline case. But in fact it contrasts sharply with Basque, because of its similarity at a syntactic level to Croatian and Bosnian. Some experts consider these to be a single, polycentric, language (Kordić, 2010; Corbett and Browne, 2018). These South Slavic languages could benefit from transfer learning, leading to an effective data size in Serbian closer to one billion tokens. No such transfer learning would be possible in Basque, an "isolate" lacking any family connection to other languages (Pereltsvaig, 2021a, page 14).

Both Gemini Pro and PaLM are trained on data that contains sentences from many human languages—a setup that favors the learning of universals, compared to monolingual training regimes. Even in such a favourable setup, they have not learned that the FOFC is a universal property that applies to all languages.

5.2 A model size threshold

Wei et al. (2022) have argued that some emergent properties start to appear only when the model size gets sufficiently large. To test if FOFC learning is affected by model size we have computed the same violation penalties for different PaLM model sizes with 8 billion parameters, 64 billion and half a trillion parameters shown in Figure 5 in the Appendix. Larger models do slightly better,

5.3 More human-like learning

If scale alone does not suffice, then what would help an LLM to learn something like the FOFC in all languages? One possibility is curriculum learning in the sense of Bengio et al. (2009, §6). Papadimitriou and Jurafsky (2023) find that pretraining on crossing dependencies helps GPT-2 when it comes to Basque, English and Japanese. Huebner et al. (2021) likewise find that training on transcribed child-directed speech leads to considerable data efficiency. Mueller and Linzen (2023) confirm that such training helps specifically on syntactic generalization of the sort tested in Section 4.

Another approach directly endows language models with inductive biases. This can be done by adding parser-like features to Transformers (Sartran et al., 2022; Murty et al., 2023; DuSell and Chiang, 2024).

6 Conclusion

In six key languages where it could have been disconfirmed, corpus analysis instead lends support to the FOFC, a candidate universal. A targeted syntactic analysis with Gemini Pro and PaLM, modern multilingual language models, indicates that the FOFC was only learned in cases where the model is exposed to developmentally-implausible amounts of training data. In Basque, where the training data size was probably closer to developmental plausibility, Gemini Pro and PaLM fail to reliably learn the constraint. So while predicting successor words can ultimately serve to learn a candidate universal, at human-scale the stimulus seems to be too poor. LLMs have learned the FOFC in some subset of languages, but they have not learned that it is a universal constraint, applicable to all human languages. Our ablation study with PaLM suggests that increasing model size helps, but not enough to change the pattern. To achieve human-level performance with language universals, it may be necessary to build in some form of inductive bias.

Limitations

This study considered just 6 human languages. It could be that the conclusion does not generalize beyond those languages. It only evaluates a few large language models. Via alternative architectures, different curricula or other training objectives it may be possible to transcend this limit and demonstrate that human-level performance is achievable via successor-word prediction within a human-sized

data budget. We have worked with the non-fine-tuned versions of the models that were trained only on maximizing text likelihood. It is imaginable that fine-tuning LLMs for syntactic universals would improve their understanding of FOFC, but the hypothesis we were interested in is whether LLMs can induce a linguistic universal without any additional supervision.

Acknowledgments

The authors would like to express heartfelt gratitude to Vera Lee-Schoenfeld at the University of Georgia (USA), Tibor Laczkó at the Károli Gáspár Reformed University (Hungary), and Ricardo Etxepare at the Research Center for Basque Language and Texts (France). Special thanks as well to Asya Pereltsvaig for her advice on Slavic, to András Kornai for timely encouragement, and to Laura Rimell for inspiring conversations about this project.

References

Giuseppe Attardi. 2015. Wikiextractor. https://github.com/attardi/wikiextractor.

Timothy Bazalgette, Ian Roberts, Michelle Sheehan, and Jenneke van der Wal. 2012. Two statistical Typological Generalisations and their Consequences. Paper presented at The Syntax of the World's Languages (SWL).

Emily M. Bender. 2013. *Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax.* Synthesis Lectures on Human Language Technologies. Springer International Publishing.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA. Association for Computing Machinery.

Theresa Biberauer. 2017. The Final-over-Final Condition and Particles. In (Sheehan et al., 2017b), chapter 9.

Theresa Biberauer, Anders Holmberg, and Ian Roberts. 2014. A syntactic universal and its consequences. *Linguistic Inquiry*, 45(2):169–225.

Vaclav Brezina. 2018. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge University Press.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. PaLM: Scaling Language Modeling with Pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Pablo Contreras Kallens, Ross Deans Kristensen-McLachlan, and Morten H. Christiansen. 2023. Large language models demonstrate the potential of statistical learning in language. *Cognitive Science*, 47(3):e13256. Letter to the Editor.

Greville Corbett and Wayles Browne. 2018. Serbo-Croat: Bosnian, Croatian, Montenegrin, Serbian. In *The World's major languages*, third edition, chapter 18, pages 339–356. Routledge.

Forrest Davis. 2022. *On the Limitations of Data: Mismatches between Neural Models of Language and Humans*. Ph.D. thesis, Cornell University.

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

Brian DuSell and David Chiang. 2024. Stack Attention: Improving the Ability of Transformers to Model Hierarchical Patterns. In *The Twelfth International Conference on Learning Representations*.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models. Unpublished technical report.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2018. SUD or surface-syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 66–74, Brussels, Belgium. Association for Computational Linguistics.

Jane B. Grimshaw. 2005. *Words and structure*. CSLI lecture notes: no. 151. CSLI Publications.

Bruno Guillaume. 2021. Graph Matching and Graph Rewriting: GREW tools for corpus exploration, maintenance and conversion. In *EACL 2021 - 16th conference of the European Chapter of the Association for Computational Linguistics*, Kiev/Online, Ukraine.

Kristina Gulordava and Paola Merlo. 2015. Diachronic trends in word order freedom and dependency length in dependency-annotated corpora of Latin and Ancient Greek. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 121–130, Uppsala, Sweden. Uppsala University, Uppsala, Sweden.

Kristina Gulordava and Paola Merlo. 2020. Computational quantitative syntax: the case of Universal 18. In Irene Vogel, editor, *Romance languages and linguistic theory 16 : selected papers from the 47th Linguistics Symposium on Romance Languages (LSRL), Newark, Delaware*, Romance Languages and Linguistic Theory: Volume 16. John Benjamins Publishing Company.

Hubert Haider. 2012. *Symmetry breaking in syntax*. Cambridge studies in linguistics: 136. Cambridge University Press.

John A. Hawkins. 2014. *Cross-linguistic variation and efficiency*. Oxford linguistics. Oxford University Press.

Anders Holmberg. 2000. Deriving OV order in Finnish. In Peter Svenonius, editor, *The derivation of VO and OV*, Linguistik aktuell: v. 31. John Benjamins Publishing CompanyBenjamins.

Anders Holmberg. 2017. The Final-over-Final Condition in a Mixed Word Order Language. In (Sheehan et al., 2017b), chapter 10.

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.

György C. Kálmán, László Kálmán, Ádám Nádasdy, and Gábor Prószéky. 1989. A magyar segédigék rendszere. *Általanos Nyelvészeti Tanulmányok*.

Hilda Koopman and Anna Szabolcsi. 2000. *Verbal Complexes*. MIT Press.

Snježana Kordić. 2010. *Jezik i nacionalizam*. Durieux (Rotulus Universitas).

Richard K. Larson. 2010. *Grammar as science*. MIT Press.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, 41(5):1202–1241.

Tal Linzen and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7(1):195–212.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Paola Merlo. 2016. Quantitative computational syntax: some initial results. *Italian Journal of Computational Linguistics*, 2(1).

Aaron Mueller and Tal Linzen. 2023. How to plant trees in language models: Data and architectural effects on the emergence of syntactic inductive biases. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11237–11252, Toronto, Canada. Association for Computational Linguistics.

Stefan Müller. 2023. *Germanic syntax*. Number 12 in Textbooks in Language Sciences. Language Science Press, Berlin.

Shikhar Murty, Pratyusha Sharma, Jacob Andreas, and Christopher Manning. 2023. Pushdown layers: Encoding recursive structure in transformer language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3233–3247, Singapore. Association for Computational Linguistics.

Frederick J. Newmeyer. 2005. *Possible and Probable Languages: A Generative Perspective on Linguistic Typology*. Oxford University Press.

Jingcheng Niu and Gerald Penn. 2020. Grammaticality and language modelling. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 110–119, Online. Association for Computational Linguistics.

Isabel Papadimitriou and Dan Jurafsky. 2023. Injecting structural hints: Using language models to study inductive biases in language learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8402–8413, Singapore. Association for Computational Linguistics.

Asya Pereltsvaig. 2021a. *Languages of the World*, third edition. Cambridge University Press.

Asya Pereltsvaig. 2021b. The OVS order in Russian: Where are the O and the V? *Journal of Slavic Linguistics*, 29(FASL 28 extra issue).

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Laurent Sartran, Samuel Barrett, Adhiguna Kuncoro, Miloš Stanojević, Phil Blunsom, and Chris Dyer. 2022. Transformer grammars: Augmenting transformer language models with syntactic inductive biases at scale. *Transactions of the Association for Computational Linguistics*, 10:1423–1439.

Michelle Sheehan. 2021. *Parameters and Linguistic Variation*, chapter 11. John Wiley & Sons, Ltd.

Michelle Sheehan, Theresa Biberauer, Ian Roberts, and Anders Holmberg. 2017a. Empirical evidence for the Final-over-Final Condition:. In (Sheehan et al., 2017b), chapter 2.

Michelle Sheehan, Theresa Biberauer, Ian Roberts, and Anders Holmberg. 2017b. *The Final-over-Final Condition: A Syntactic Universal*. MIT Press.

Jon Sprouse, Beracah Yankama, Sagar Indurkhya, Sandiway Fong, and Robert C. Berwick. 2018. Colorless green ideas do sleep furiously: gradient acceptability and the nature of the grammar. *The Linguistic Review*, 35(3):575–599.

Anatol Stefanowitsch. 2006. Negative evidence and the raw frequency fallacy. *Corpus Linguistics and Linguistic Theory*, 2(1):61–77.

Anatol Stefanowitsch. 2020. *Corpus linguistics*. Number 7 in Textbooks in Language Sciences. Language Science Press, Berlin.

Timothy Stowell. 1981. *Origins of phrase structure*. Ph.D. thesis, MIT.

Luka Terčon and Nikola Ljubešić. 2023. Classla-stanza: The next step for linguistic processing of south slavic languages.

Harry van der Hulst. 2023. *A Mind for Language: An Introduction to the Innateness Debate*. Cambridge University Press.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Arnold M. Zwicky. 1985. Heads. *Journal of Linguistics*, 21(1):1–29.

A Tree rotation script

The other necessary function, `make_head_final` is analogous but opposite.

```
def phrase_headed_by(token_number: int, government_table: dict[int, list[int]]
                     ) -> set[int]:
  # Transitive closure of the government table returning
  # all descendents of token_number.
  if token_number in government_table:
    direct = set(government_table[token_number])
    return {token_number} | direct.union(*[phrase_headed_by(x, government_table)
                                           for x in direct])
  else:
    return {token_number}

def footprint_of_phrase(token_number: int, government_table: dict[int, list[int]]
                        ) -> set[int]:
  # All indices between left-most and right-most descendent.
  phrase = phrase_headed_by(token_number,government_table)
  return set(range(min(phrase), max(phrase)))

def leftmost_in_phrase(token_number: int, government_table: dict[int, list[int]]
                       ) -> int:
  return min(phrase_headed_by(token_number,government_table))

def rightmost_in_phrase(token_number: int, government_table: dict[int, list[int]]
                        ) -> int:
  return max(phrase_headed_by(token_number,government_table))

def phrase_excluding_comp(head_number: int, comp_number: int,
                          government_table: dict[int, list[int]]) -> set[int]:
  # All words governed by head_number excluding words coverned
  # by its complement comp_number.
  return (phrase_headed_by(head_number,government_table)-
          phrase_headed_by(comp_number,government_table))

def footprint_excluding_comp(head_number: int, comp_number: int,
                             government_table: dict[int, list[int]]) -> set[int]:
  return (footprint_of_phrase(head_number,government_table)-
          footprint_of_phrase(comp_number,government_table))

def leftmost_in_phrase_excluding_comp(head_number: int, comp_number: int,
                                      government_table: dict[int, list[int]]
                                      ) -> int:
  return min(phrase_excluding_comp(head_number, comp_number, government_table))

def rightmost_in_phrase_excluding_comp(head_number: int, comp_number: int,
                                       government_table: dict[int, list[int]]
                                       ) -> int:
  return max(phrase_excluding_comp(head_number, comp_number, government_table))

def make_head_initial(h: int, d: int, heads: list[int]) -> list[int]:
  """Transform tree, as defined by dependent -> head mapping in variable heads,
     into a new tree where a head-final dependency h->d is converted
     into a head-initial one.
  """
  n = len(heads) # includes the root at position 0

  # invert the head relation
  # for every node that has a head, what are indexes of nodes that share that head?
  governs: dict[int, list[int]] = toolz.groupby(lambda i: heads[i], range(n))

  # shift the dependent rightwards across the length of the entire governing phrase
  shift_d_right = rightmost_in_phrase(h,governs) - rightmost_in_phrase(d,governs)

  # shift the governing phrase left across the length of the entire complement
  shift_h_left = (leftmost_in_phrase_excluding_comp(h,d,governs)
                  - leftmost_in_phrase(d,governs))

  # make a map from old positions to new positions
```

```
  new_heads = []
  for i in range(n):
    if i in  footprint_of_phrase(d,governs):
      new_heads.append(i+shift_d_right)
    elif i in footprint_excluding_comp(h,d,governs):
      new_heads.append(i-shift_h_left)
    else:
      new_heads.append(i)   # no change
  return new_heads
```

B Example corpus query

This appendix documents the query that was used to find cases of the FOFC-violating configuration 1d (i.e. the gray cell in Figure 1). The query is expressed here in the Grew notation (Guillaume, 2021), which is particularly clear. The actual search was performed by a lower-level implementation in python.

In Hungarian, rather than relying on the AUX part of speech tag, we instead use an explicit list of "auxiliaries" (Kálmán et al., 1989, cited in Koopman and Szabolcsi). This is implemented in the disjunctive regular expression specification on the lemma feature of AUX.

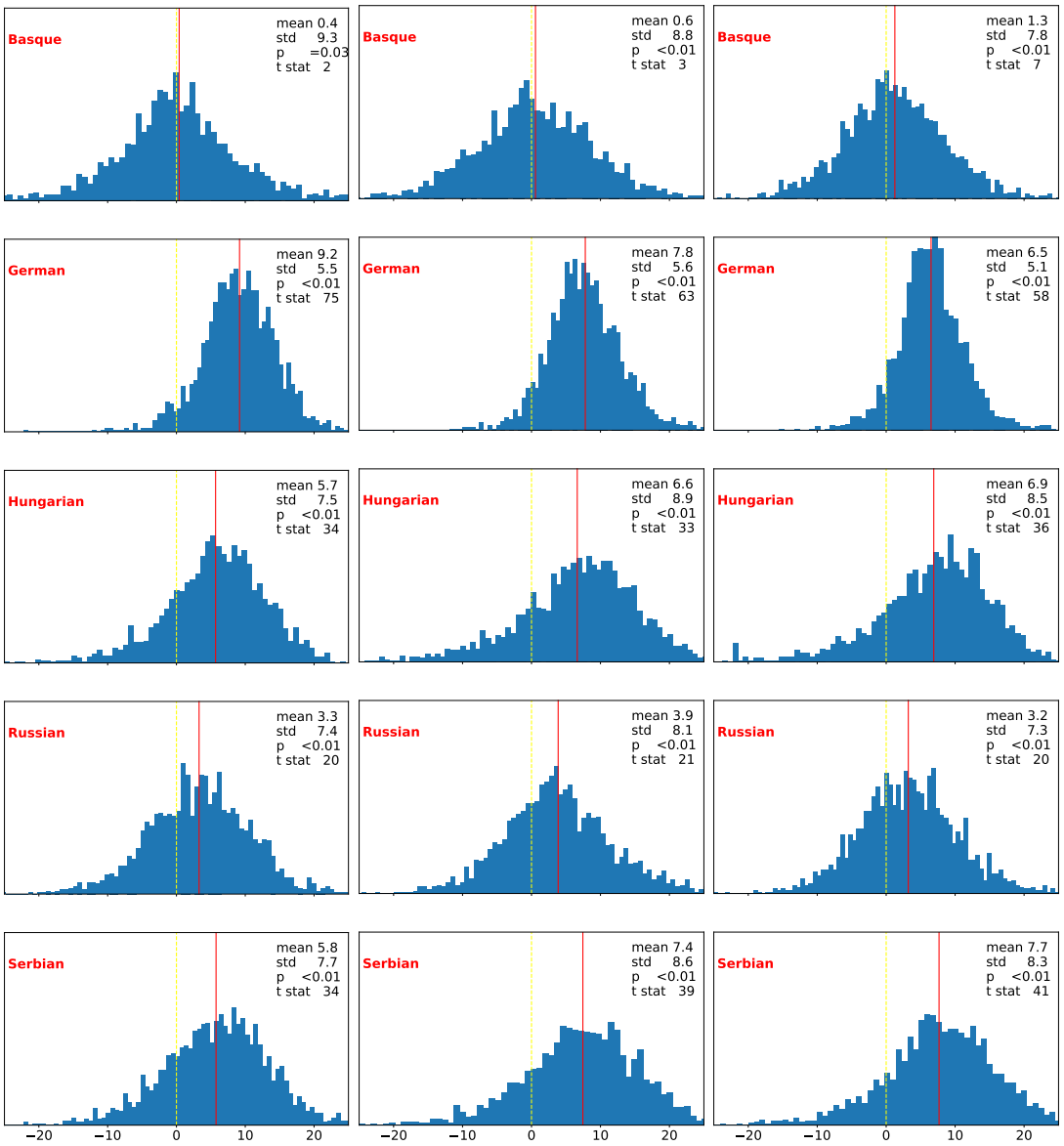
```
pattern {
  AUX [lemma=re"fog\|lehet\|szokott\|szokás\|tetszik\|szabad\|szeretne\|
                kell\|akar\|talál\|bír\|tud\|kezd\|kíván\|mer\|óhajt\|
                próbál\|szándékozik"];
  V [upos=VERB];
  AUX -[comp:aux]-> V;

  O [upos=NOUN];
  V -[comp:obj]-> O;

  V << AUX; % AUX further to the right
  O << AUX; % than either element of its complement VP

 V <O; % head initial VP downstairs
}
% no punct between leftmost element and rightmost
without{
 P[upos=PUNCT];
 V << P;
 P << AUX
}
% no conjunction between leftmost and rightmost
without{
 CC[upos=CCONJ];
 V << CC;
 CC << AUX
}
```

C Additional constraints applied in Slavic languages

Corpus search of Russian and Serbian applied two additional constraints in addition to those discussed in Section 3 of the main text. These constraints, listed below, are intended to exclude cases of VP-fronting.

- the lexical verb V may not occur sentence-initially
- nor may it immediately follow a conjunction (SCONJ or CCONJ), either at the beginning of the sentence or immediately after a punctuation symbol (PUNCT)

As Pereltsvaig (2021b) persuasively argues, such VP-fronting is A-bar movement. This would be analogous to VP topicalization in German or contrastive topicalization of both verb and object in Hungarian, neither of which is relevant to the FOFC as applied to two-phrase configurations of {Aux, V}. The invocation of the PUNCT tag in the second bulleted constraint is intended to capture cases where punctuation signals a clause boundary.



(a) PaLM 8B (b) PaLM 64B (c) PaLM 540B

Figure 5: The results with different sizes of PaLM model. As before, x-axis represents the difference in log-probability of FOFC abiding and FOFC breaking word orders, while y-axis represents the ratio of samples that have score in that range.