COPiL CAMBRIDGE OCCASIONAL PAPERS in LINGUISTICS
Volume 15, Article 3: pp. 55–110, 2023 | ISSN: 2050-5949

On the Potential for Maximising Minimal Means in Transformer Language Models: A Dynamical Systems Perspective[*]

Suchir Salhan
University of Cambridge

Abstract Computational linguists can utilise the insights of neo-emergent linguistic models, an approach to grammar construction that relies heavily on domain-general inductive biases, to address extant challenges associated with the syntactic and typological capabilities of state-of-the-art Transformer-based Language Models (LMs), which underpin systems like Google Translate and ChatGPT. I first offer a synthesis of the inductive biases of Transformer-based LMs that are reminiscent of two properties emphasised in a neo-emergent model called 'Maximise Minimal Means' (MMM) (Biberauer 2011: *et seq*). Subsequently, I undertake an analysis of the structural generalisation capabilities of Transformer-based LMs through a creative probing case-study of subject-verb agreement, which indicates that these models are unable to perform the crucial NO > ALL > SOME learning dynamic associated with MMM. In light of these empirical findings, I offer a theoretical argument about how MMM and associated Dynamical Systems Theory (Bosch 2022, 2023) can be viewed as a linguistically motivated goal– as proposed by Emerson (2020b) – for Transformer LMs. I propose that the predictions of this neo-emergentist approach translate into theoretical principles and practical recommendations to improve the syntactic capabilities of Transformer-based LMs in a typologically-consistent manner. This perspective can stimulate a productive interdisciplinary discussion on how linguistic theory can help engineer LMs with better syntactic and typological capabilities.

1 Introduction

Inductive Biases play a fundamental role in grammar construction and underpin the structural-generalisation capabilities of Language Models in computational linguistics. The inductive biases of a family of state-of-the-art neural monolingual and multilingual *Transformer Language Models (LMs)* (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser & Polosukhin 2017) have been widely studied.

[*] I am hugely indebted to my supervisor, Dr Andrew Caines (Computer Science & Technology) for his support and encouragement throughout this dissertation. Many thanks to Dr Theresa Biberauer, Núria Bosch-Masip, and Chris Davis for their guidance and advice at various stages. I would also like to thank Prof Paula Buttery, Prof Nigel Collier, Fangyu Liu and Dr Li Nguyen for their supervision of various research projects in TAL and the NLIP Group (Computer Science) and the comments and suggestions from the audience of my SyntaxLab Talk on 14/02/2023 that have greatly shaped thoughts and ideas. Any errors are, of course, my own.

Transformers, which underpin Google Translate's Neural-Machine-Translation system and OpenAI's Chat-GPT, face three challenges: few-shot learning, Transfer Learning from high-to-low-resource languages, and syntactic generalisation. I argue that these issues can be approached using neo-emergent accounts of language acquisition and typological variation, where learners recruit domain-general inductive biases to steer them towards salient properties of the input. Neo-emergentism redistributes the division-of-labour in grammar construction towards emergent representations (Wiltschko 2014, 2021, Ramchand & Svenonius 2014).

Specifically, I motivate the relevance of one domain-general inductive bias called *Maximise Minimal Means (MMM)* (Biberauer 2011, 2015, 2017, 2019a,b, van der Wal 2022, Bosch 2022, 2023). MMM is a two-step meta-learning algorithm for data-efficient grammar construction, driving the learner towards emergent syntactic acquisition in a manner that underpins typological patterns. I offer an exposition to Transformer-based LMs (which are firmly rooted in the Firthian adage that 'you shall know the meaning of a word by the company that it keeps' Firth 1957) and MMM. I propose a previously unnoticed convergence in section 2 in how Transformers and learners recruit their inductive biases. Emergent syntax is supported by pre-training dynamics in the Transformer that approximate the dynamics of emergence predicted in MMM. State-of-the-art work on meta-learning and structural generalisation is also reminiscent of MMM-approaches.

This raises two research questions – formulated in section 3. Many computational linguists are focused on 'climbing the right hills' (Bender & Koller 2020) using top-down (often semantico-centric) goals. This aims to improve structural generalisation and transparency alongside 'bottom-up' intrinsic evaluation metrics, like perplexity, to engineer better-performing LMs (Manning 2015).

Approaching Transformer-based LMs from a syntactico-centric perspective, I analyse their emergent capacity to learn English subject-verb agreement using two 'creative' probing strategies in section 4. Transformers face a fundamental limitation in structural-generalisation.

Under the spirit of this 'top-down' approach to Language Modelling espoused by (Emerson 2020b), I propose in section 5 that MMM can guide the improvement of the structural generalisation capabilities of LMs down a 'syntatically-motivated cline'. This allows us to:

i. Develop linguistically-informed syntactic complexity metrics. I propose one metric, called *sensitivity*, which allows us to assess whether pre-training and transfer learning techniques in Transformers replicate the emergent dynamics of MMM. I highlight the limitations of commonly-used metrics.

ii. Formalise upper bounds on the potential of existing techniques that augment the Transformer to distil and transfer syntactic structure across languages. I present a *typological upper-bound on syntactic transfer learning*.

iii. Engineer machine learning techniques to replicate the dynamics of syntactic emergence, as predicted by MMM while retaining the core advantages of the Transformer architecture in downstream tasks. I suggest that a combination

> of pre-training LMs using *child corpora of increasing linguistic complexity* and *distilling emergent syntax* from contextual embeddings *should* lead to better structural generalisation. I hope to develop machine learning techniques in the future to implement this prediction.

This analysis of emergent syntax is significant as it makes a theoretically-motivated shift to view the Transformer more abstractly to ascertain its capability to replicate the learning dynamics that derive typological distributions– useful in the impetus towards building 'fair-NLP' solutions that can be equitably applied across languages and cultures. This 'MMM-approach' provides computational linguistics with a narrow roadmap of exactly *how* existing machine-learning techniques lead to human-like syntactic emergence. It prompts the development of better techniques that algorithmically realise and practically-implement theoretical models of what is possible and probable in natural language in a connectionist architecture.

2 Background and Literature Review

After offering an exposition of the relevant background on the Transformer architecture and associated model variants and the acquisitional and typological proposals of the MMM model, I provide a novel characterisation of the convergence of work on meta-learning in theoretical and computational linguistics.

2.1 Transformer Language Models

Transformer LMs are the contemporary state-of-the-art architecture used in computational linguistics, succeeding previous neural LMs, like LSTMs (Long Short Term Memory) and RNNs (Recurrent Neural Networks). A Language Model is a probability distribution over a sequence of words. Classical n-gram LMs have a training objective of predicting the next word w_{n+1} in a sequence given n words of the preceding context, $P(w_{n+1}|w_1, \ldots, w_n)$. Neural network LMs use *embeddings*, which are continuous (typically) vector representations of words to make next-word predictions. Other embedding representations, such as tensors (Coecke, Sadrzadeh & Clark 2010, Gong, Bhat & Viswanath 2018), can be used.

The introduction of low-dimensional static embeddings by Mikolov, Chen, Corrado & Dean (2013a), Mikolov, Sutskever, Chen, Corrado & Dean (2013b) led to the adoption of self-supervised representation learning. These do not require any human-annotated labels and can be created from entirely unlabelled datasets– a lightweight process for generating text-representations that can be used for downstream tasks. Static embeddings are generated by minimising a loss function to predict a target word w_i from a fixed input vocabulary $|V|$ for a given context window of words preceding and following the target. Reversing this training objective in a Skip-Gram model allows models to formulate a semantic conceptual space from which vector additive composition can approximate a natural language phrase – famously the embedding for QUEEN is the embedding that maximises cosine similarity with KING $-$ MAN $+$ WOMAN (Mikolov et al. 2013a: 4).

The key architectural innovation of Transformers is the use of attention. Bahdanau, Cho & Bengio (2014) introduced an attention mechanism, inspired by techniques on 'learning to align' in sequence-to-sequence (seq2seq) networks used in transduction tasks like machine translation where the embedding of an 'encoded' source sequence is mapped to a 'decoded' output sequence.

Vaswani et al. (2017) introduce the Transformer Language Model, illustrated in Figure 1, where its encoder (left), which learns contextualised representations, and decoder (right) are comprised of six stacked multi-attention heads, each fully connected by a feedforward neural network. This relates different positions of a text sequence to compute a representation by 'attending' to parts of the input sequence that the model seems relevant to current predictions. Transformers rely solely on the attention mechanism to draw global long-range dependencies between input and output sequences, eschewing the use of recurrence in previous RNN-based architectures. Embeddings are also used in Transfer Learning: the linguistic knowledge encoded in embeddings is transferred to help the Transformer improve performance on related tasks or languages.

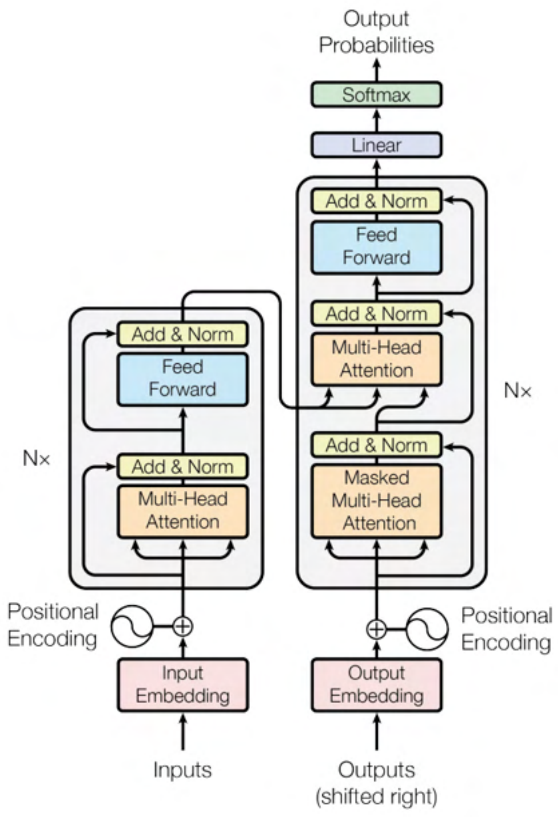


Figure 1 The Transformer architecture. Figure taken from Vaswani et al. (2017).

2.1.1 Transformer Self-Attention Mechanism

Following Hahn (2020: 158-159), I outline how the Transformer self-attention mechanism works. Given some finite alphabet $|V|$, there is an input $x = x_1, x_2, \ldots, x_n$ where all $x_i \in |V|$ and x_n is an end-of-sequence $[EOS]$ symbol. The Transformer encodes this input-string into a sequence of input embeddings $v_1, v_2, \ldots, v_n$ using some embedding map $V \rightarrow \mathbb{R}^k$. The Transformer has a sequence of positional embeddings $p_1, p_2, \ldots, p_i \in \mathbb{R}^k$ that are independent of input x. They are either computed using a pre-defined scheme or learnt for each position occurring in the training data. Using either addition or concatenation, the input and positional embeddings are combined into the vectors of Layer 0 of the Transformer, $y_i^{(0)} = f(v_i, p_i)$.

Each layer of the Transformer has a set of H attention heads which combine the point-wise 'activation' $y_i^{(k)}$ for position i at layer k. Vaswani et al. (2017: 5) implements attention scores by linearly transforming $y_i^{(k)}$ into Query and Key vectors. The Attention Score, $a_{i,j}^{(k,h)}$, combines the activations from previous levels, $a_{i,j}^{(k,h)} = f_{k,h}^{att}(y_i(k-1), y_j(k-1))$.

The activation of a head is computed by either attending to positions with the maximum attention value in the hard attention variant of the Transformer or applying a softmax function to the scaled dot-product of the query and key vectors, Q and K in the soft attention variant:

(1) $$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

The most 'confident' attention-heads play linguistically-interpretable roles that clearly focus on 'syntactically-relevant' positions in a sentence, attending to adjacent words and tracking syntactic dependencies (Voita, Talbot, Moiseev, Sennrich & Titov 2019).

2.1.2 Transformer-based Language Models

Within the contemporary ecology of state-of-the-art Language Models, the 'vanilla' Transformer architecture has been adapted using different pre-training objectives and augmented using sub-word tokenisation; increasing model size and extending the Transformer to multilingual and multimodal settings. I introduce the Transformer-based models that are used in section 4.

Models, like Bidirectional Encoder Representations for Transformers (BERT), use a *masked language modelling (MLM)* training objective, where final hidden vectors are randomly masked following a procedure inspired by cloze-style tasks used in psycholinguistics. Devlin, Chang, Lee & Toutanova (2019), building on ELMO (Peters, Neumann, Iyyer, Gardner, Clark, Lee & Zettlemoyer 2018), develop BERT's text representations used in downstream tasks by extracting context-sensitive word embeddings from bidirectional left-to-right and right-to-left language models. The BERT architecture is fed input sentences, which are encoded using pre- trained token embeddings trained using MLM and concatenated with positional embeddings and [SEP] tokens that associate tokens of the same sentence. The MLM objective enables deep bidirectional pre-training in BERT, by masking subword tokens used as input vocabulary at random and then training BERT to predict these tokens. BERT also uses a next-sentence-prediction training objective to jointly pre-train text-pair representations. However, RoBERTa (Liu, Ott, Goyal, Du, Joshi, Chen, Levy, Lewis, Zettlemoyer & Stoyanov 2019) is a variant of the BERT model, trained on a much larger dataset with a more effective training procedure. RoBERTa removes the next-sentence-prediction objective, which was found to be less important for model performance. It uses a dynamic masking technique during training, which helps the model learn more robust and generalisable representations of words. Transformer-based Language Models that use a MLM training objective differ variously: one important dimension is model size, where $BERT_{BASE}$ has 12 layers, while $BERT_{LARGE}$ has 24 layers.

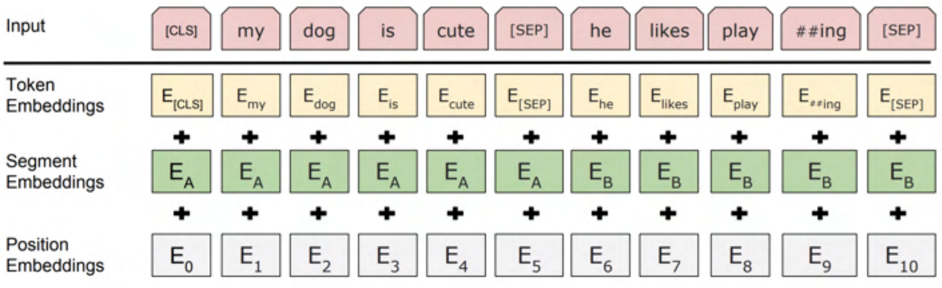


Figure 2 The BERT architecture. Figure from Devlin et al. (2019).

ELECTRA (Clark, Luong, Le & Manning 2020) uses a more sample-efficient pre-training objective called *replaced token detection*, which corrupts a percentage of the input by replacing target tokens with plausible alternatives that are sampled from a small generator network. The objective of a discriminator model is to predict whether each token in the input is corrupted or not. As the task is defined over the entire input rather than a percentage of masked tokens, Clark et al. (2020) find that this approach can 'substantially outperform the ones learned by BERT given the same model size, data, and compute'.

Large gains in Natural Language Understanding tasks have been made using Generative Pre-Training of Transformer LM on a diverse corpus of unlabeled text, followed by supervised 'discriminative' fine-tuning on specific downstream tasks. The GPT family (Radford, Narasimhan, Salimans, Sutskever et al. 2018) uses a standard *'autoregressive' language modelling objective* for next-word prediction to maximise the log-likelihood of $P(w_i|w_{i-k},\ldots,w_{i-1},\Theta)$where k is the size of the context window for the unsupervised pre-training of the 'vanilla' Transformer decoder (using unidirectional left-to-right self-attention). The decoder produces an output distribution over the target tokens. The GPT-family are Large LMs, with state-of-the-art variants seeing a massive explosion in training corpus size and the number of configurational model 'parameters' which estimate statistical patterns– GPT-2 and GPT-3 (Brown, Mann, Ryder, Subbiah, Kaplan, Dhariwal, Neelakantan, Shyam, Sastry, Askell et al. 2020) have 1.5 and 175 billion parameters respectively. GPT-3 uses the same architecture as GPT-2, with the exception that it uses alternating dense and locally banded sparse attention patterns in the layers of the transformer, similar to the Sparse Transformer. Brown et al. (2020) find that scaling up the size of LMs can lead to much better 'on-the-fly' few-shot learning performance in task-agnostic settings.

LMs also differ in their subword tokenisation algorithms: BERT uses WordPiece tokens (Schuster & Nakajima 2012), while GPT-2 and GPT-3 use Byte-Pair Encoding (Sennrich, Haddow & Birch 2016).

Transformer-XL (Dai, Yang, Yang, Carbonell, Le & Salakhutdinov 2019) is an autoregressive English LM whose training objective is similar conceptually to GPT-2's; but it has a much longer effective context to enable learning longer-term dependencies using a segment-level recurrence scheme that allows the Transformer to fully-exploit its optimisation advantage on the 'vanishing gradients' problem– effectively preventing model weights to be updated during pre-training. XLNet (Yang, Dai, Yang, Carbonell, Salakhutdinov & Le 2019) is an English LM which proceeds through various word order permutations of the input tokens during training, and which uses a distinct attention masking mechanism as well; during testing, it proceeds autoregressively through the input similar to the other two models.

Recent research has focussed on the development of *multilingual sentence encoders*, like multilingual BERT (mBERT), which are trained in as many as 104 languages to jointly train a shared embedding representation space in a multilingual encoder. These enable immediate (or zero-shot) cross-lingual transfer (see section 5.3).

There are also various techniques that are aimed to improve the few-shot learning potential of Transformer-based models in fine-tuning, such as Pattern-Exploiting Training (Schick & Schütze 2021): a semi-supervised training procedure that reformulates input into cloze-style templates to enable pre-trained Transformer-based LMs to predict continuations to solve downstream tasks like sentiment analysis. The Transformer architecture has been integrated into multimodal architectures in models like CLIP (Radford, Kim, Hallacy, Ramesh, Goh, Agarwal, Sastry, Askell, Mishkin, Clark et al. 2021), which integrate image information with unimodal text-representations from GPT-2.

2.2 Maximise Minimal Means

Neo-emergent models redistribute the division of labour between the Three Factor Model of Chomsky (2005). Modifying the ontological apparatus of the generative enterprise, they endow the Language Faculty with universal symbolic operations and components that are conceptually-necessary, but still maintain symbolic computation unlike usage-based approaches and Construction Grammar. In addition to Factor 1– a maximally-poor UG with operations like Merge and Agree relying on a set of formal features, FF– and Factor 2 (primary linguistic data provided to the learner), MMM is a *domain-general third-factor principle recruited for language-specific computation.*

2.2.1 MMM as a Third Factor

Neo-emergent systems are driven to 'maximise minimal means' as follows. The initial state of a system has NO postulated representations, however, learners exhibit sensitivity to salient contrasts. This triggers the postulation of a feature in a maximally-wide domain (ALL), before restricting its application to a more context-sensitive domain upon receiving further input (SOME). Following cyclic NO > ALL > SOME patterns of over-extension and retraction leads to an expectation that later knowledge will 'piggyback' on and be shaped by pre-existing knowledge. The MMM inductive bias drives learners to follow an acquisitional-pathway that yields so-called *'Goldilocks Effects'*, where learners systematically attend to salient information in the input to avoid overfitting; combat noise; and ignore some of the inherent complexity in the environment. MMM+Goldilocks yields a strategy choice that is adaptive and 'good-enough' based on the problem's characteristics (Biberauer & Bosch 2021).

2.2.2 MMM in the Language Faculty

While Chomsky (2000) assumes that grammars make a 'one-time selection' of formal features, FF, used contrastively in narrow syntax in the Language Faculty, MMM make a significant departing assumption that features are only acquired by the learner if the PLD necessitates so. The first condition of the MMM model is that learners exhibit sensitivity to many-to-one/one-to-many form-meaning mappings (*Systematic Departures from Saussurean Arbitrariness*) (Biberauer 2019b). Subject-verb agreement is an example of a many-to-one departure from Saussurean arbitrariness, which we discuss in section 4. Agreement is a conventionalised grammatically-regulated dependency marked on $\text{NP}_{\text{subject}}$ and V. When a learner encounters subject-verb agreement, they are forced to postulate ϕ-(person/number/gender) FFs used contrastively in steady-state agreement computation. Emergent syntax is triggered by the MMM inductive-bias towards salient cues of 'higher degrees' of grammatically-regulated arbitrariness in the input.

Once FFs are postulated, MMM guides the learner to follow a domain-general NO > ALL > SOME geometry, which is manifested as a maximally-wide extension of FF to 'relevant' syntactic heads (as specified by universal computation in narrow syntax), before retracting FF to more context-specific head distribution to account for more fine-grained distributional patterns in PLD. NO > ALL > SOME underpin the distributional patterns of syntactic typology as the locus of Parametric variation is attributable to differences in FF-specifications associated with (functional) syntactic heads (Baker 2008). The MMM contextual restriction is illustrated in Figure 3.

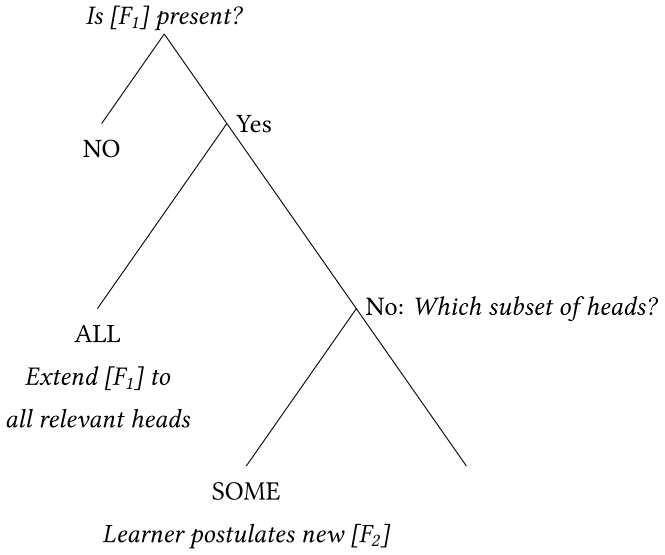


Figure 3 Departures from Saussurean Arbitrariness trigger MMM contextual restrictions.

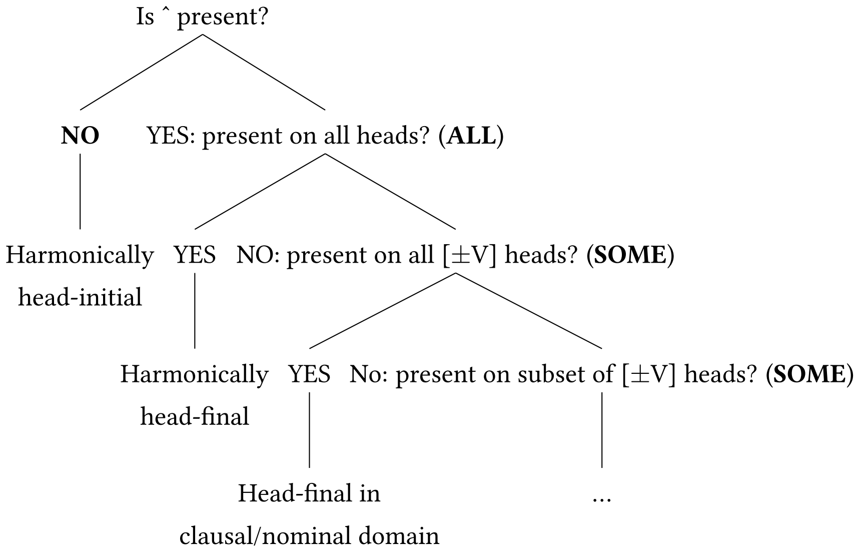

Is ˆ present?
NO
YES: present on all heads? (**ALL**)
Harmonically
head-initial
YES
NO: present on all [±V] heads? (**SOME**)
Harmonically
head-final
YES
No: present on subset of [±V] heads? (**SOME**)
Head-final in
clausal/nominal domain
...

Figure 4 Parameter Hierarchy for Word Order Typology (Roberts 2012).

Concretely, word order typology 'falls-out' by initially postulating a feature-diacritic $\wedge$, triggered by very-early prosodic-bootstrapping. This derives VO and OV harmonic word-order (NO > ALL). Subsequent disharmonic orders are derived by restricting the distribution of $\wedge$ to a subset of heads based on further input (SOME). The MMM-dynamic derives emergent Parameter Hierarchies, a device that syntacticians have proposed to account for typological variation (van der Wal 2022: i.a.).

2.3 Meta-Learning and Inductive Biases

I review the predictions of Dynamical Systems Theory (Bosch 2022, 2023) that extends MMM and propose the relevance of a complexity metric called sensitivity for emergent syntax in Transformers. I provide a novel synthesis of inductive biases and meta-learning in these two approaches. This isolates two shared inductive biases that both Transformers and MMM-learners recruit in emergent syntax.

2.3.1 Dynamical Systems Theory

Neo-emergent *Dynamical Systems Theory (DST)* reconciles how learners postulate emergent symbolic-representations despite external environmental perturbations caused by the dynamics of areal/contact-induced/diachronic change. The Language Faculty, schematised in Figure 5, has an acquisitional dynamical system $< T, X, \Phi >$, composed of a set of times T, a state space X and a set of transition functions $\Phi : X \times T \to T$ that interacts with a system of symbolic computation, alongside an intermediate Conceptual Space where (semantic) concepts are represented as convex sets.

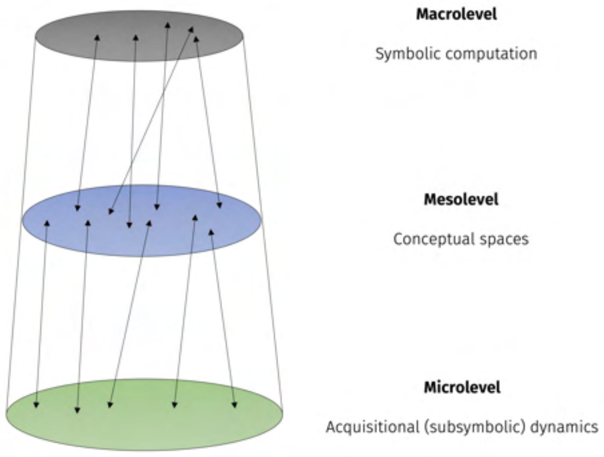


Figure 5 The Language Faculty as conceived in Dynamical Systems Theory. Figure from Bosch (2022: 17).

Universal Grammar specifies the flow (rather than the hard-wired substance) of emergent representations in the acquisitional dynamical system. MMM and Goldilocks Effects inductive biases are jointly-encoded as 'attractors' in the dynamical system, steering learners to follow MMM dynamics as a way to adaptively mutate the grammar to handle noisy perturbations in the input. The acquisitional dynamical system uses contrastive cues as a 'control parameter' to drive the successive re-organisation of the system especially to early input in the initial stages of grammar construction.

DST has two relevant predictions: (1) the inductive biases underpinning emergent syntax can be analysed using dynamical systems and (2) typological variation 'falls-out' from the greater initial learning potential of the dynamical system to early input. DST addresses three extant problems also addressed in work that augments the performance of the 'vanilla' Transformer architecture to support the emergence of contextual embeddings in a typologically-consistent manner.

i. *Emergent Neuro-Symbolic Representation:* DST proposes that learners distil emergent symbolic representations from continuous learning dynamics. Motivated by the limitations of LMs to model truth functions, semantic compositionality and quantifier scope, Functional Distributional Semantics augmented continuous vector-based embeddings with compositional semantic computation (Emerson 2018, 2020a). Both converge on neuro-symbolic representations.

ii. *(Domain-General) Meta-Learning:* MMM is a domain-general/'third-factor' meta-learning (or *learning-how-to-learn*) procedure. The Transformer is a domain-general neural architecture – applied in non-language-specific contexts, like computer vision (Dosovitskiy, Beyer, Kolesnikov, Weissenborn, Zhai, Unterthiner, Dehghani, Minderer, Heigold, Gelly et al. 2020). Meta-learning algorithms that augment the Transformer architecture aim to improve the compositional generalisation of Transformer LMs (Conklin, Wang, Smith & Titov 2021) and in cross-lingual transfer (Ponti 2021). Both recruit domain-general meta-learning procedure for language-specific computation

iii. *Typology:* The MMM meta-learning algorithm steers the patterns of emergence in the dynamical system that underpins typological distributions. This has potentially relevant theoretical consequences for understanding the theoretical limits of Transfer-based approaches.

I now motivate a complexity metric called sensitivity which elucidates how Transformers recruit their domain-general inductive biases to exhibit two DST properties: (1) sensitivity to initial conditions and (2) emergent complexity-driven curriculum learning.

2.3.2 *Sensitivity and Self-Attention*

Sensitivity is a complexity metric that measures how likely a function value changes based on the number of bits flipped in the input.[1] Functions with low sensitivity have a low Kolgomorov complexity (the information in x is defined by the length of the shortest program encoded in binary bits that outputs the string x)– they depend on fewer bits and are determined by a smaller number of input coordinates. Functions of higher sensitivity are more complex since the function value can be changed by any subset of bits (Novak, Bahri, Abolafia, Pennington & Sohl-Dickstein 2018).

The Transformer Self-Attention mechanism can be modelled as a dynamical system of interacting particles. The numerical solution of the Ordinary Differential Equation that models the Transformer at a time t is related to the set of input embeddings fed into the Transformer and the Key/Query vectors in the multi-head self-attention mechanism. [2]

Inoue, Ohara, Kuniyoshi & Nakajima (2022) further illustrate that the pre-training procedures used in a smaller version of BERT, ALBERT (Lan, Chen, Goodman, Gimpel, Sharma & Soricut 2019), can be viewed as a dynamic trajectory along a discrete-time dynamical system from its initial state as a randomly initialised network. When input is provided to ALBERT, the encoder is in a transient state until it reaches a certain state. They characterised smooth and stable short-term training, where embeddings began to synchronise due to the influence of the self-attention mechanism, indicating natural language understanding tasks can be handled by a discrete dynamical system.

2.3.3 *How does Sensitivity allow us to understand emergent MMM-like properties in Transformers?*

We now apply the definition of sensitivity and the Dynamical Systems interpretation of the self-attention mechanism to highlight two inductive biases of the Transformer.

The Transformer Self-Attention mechanism exhibits a simplicity bias that is reminiscent of MMM-like inductive bias attributed to the acquisitional dynamical system in DST. Transformers exhibit a *simplicity bias towards sparse boolean functions with low sensitivity*. In these conditions, they exhibit the capability to generalise near perfectly even in the presence of noisy labels, unlike LSTMs which overfit and achieve poorer generalisation accuracy (Bhattamishra, Patel, Kanade & Blunsom 2022). The existence of a simplicity bias in Transformers is significant from an MMM-perspective, which also assumes that human learners have a bias towards simpler inputs which drive the flow of emergent syntax. Edelman, Goel, Kakade & Zhang (2022) implicate *sparse features in the emergence of representations that are reminiscent of the hand-crafted syntactic features.*[3] A single self-attention head

[1] See Kalemaj (2020) for a formal definition.

[2] See Dutta, Gautam, Chakrabarti & Chakraborty (2021: 3-4) for a proof of this result.

[3] Edelman et al. (2022: 7) proves that the self-attention mechanism can represent all sparse Boolean functions for any size-k subset, which allows the Transformer to learn sparse interactions sample-efficiently

in Transformer networks creates sparse variables of the input sequence context, with sample complexity scaling only logarithmically with the context length T. As predicted by DST, the emergent structural generalisation capabilities of the Transformer architecture are highly sensitive to the initial conditions of the training data.

MMM-learners tend to pick analyses with the shortest description length of the linguistic input. The application of curriculum learning to Transformers allows models to selectively attend to information at varying levels of granularity– although, unlike MMM-based curricula, these typically take the form of ordering a sequence of tasks or sampling a dataset according to a pre-specified order (Surkov, Mosin & Yamshchikov 2022).

However, Transformers implicitly proceed in an emergent curriculum from low-to-high sensitivity. Syntactic Probing investigates representations associated with syntactic dependencies by developing small supervised models called probes that map from model representations to some phenomenon that those representations are expected to encode. The *Structural Probe* decodes dependency parse trees from the self-attention mechanism by finding a linear transformation under which two words' distance in their dependency parse is approximated by the squared distance between their model representation vectors under a linear transformation that defines a syntactic subspace (Hewitt & Manning 2019). This is done by recovering a unidirectional graph that maps contextualised embeddings into the syntactic subspace with a distance measure optimised to the number of edge spans between two words in a dependency graph (Hewitt & Manning 2019: 4130). It encodes which word is governed by other words and each word's proximity to every other word in the syntax tree.

DEPPROBE is an extension of the Structural Probe that extracts fully-labelled, directed dependency trees from the Transformer self-attention mechanism' (Müller-Eberstein, van der Goot & Plank 2022a). By combining the distance measure of the Structural Probe with a relational probe that learns the probability of a word being classified into a dependency relation, the dependency graph is rooted by the word that has the highest probability of being a root, and the iterates across all words until they are covered in an edge drawn from its head based on the distance measure of the Structural Probe. This creates a linear probe that extracts a fully labelled and directed dependency parse from contextualised embeddings."

Probing has established that full trees can be decoded above baseline accuracy from single attention heads, and that individual relations are often tracked by the same heads across languages (Ravishankar, Kulmizev, Abdou, Søgaard & Nivre 2021, Limisiewicz, Mareček & Rosa 2020, Chi, Hewitt & Manning 2020), as illustrated in Figure 6.

Transformers filter linguistic information along different linguistic timescales, such as words, phrases and sentences, and associate different linguistic tasks with these bandwidths. Building on Tamkin, Jurafsky & Goodman (2020), Müller-Eberstein, van der Goot & Plank (2022b) develop a technique called Spectral Probing, which uncovers emergent curricula in Transformers. After decomposing a sequence of embeddings into a composite frequency wave using Discrete Cosine Transform,

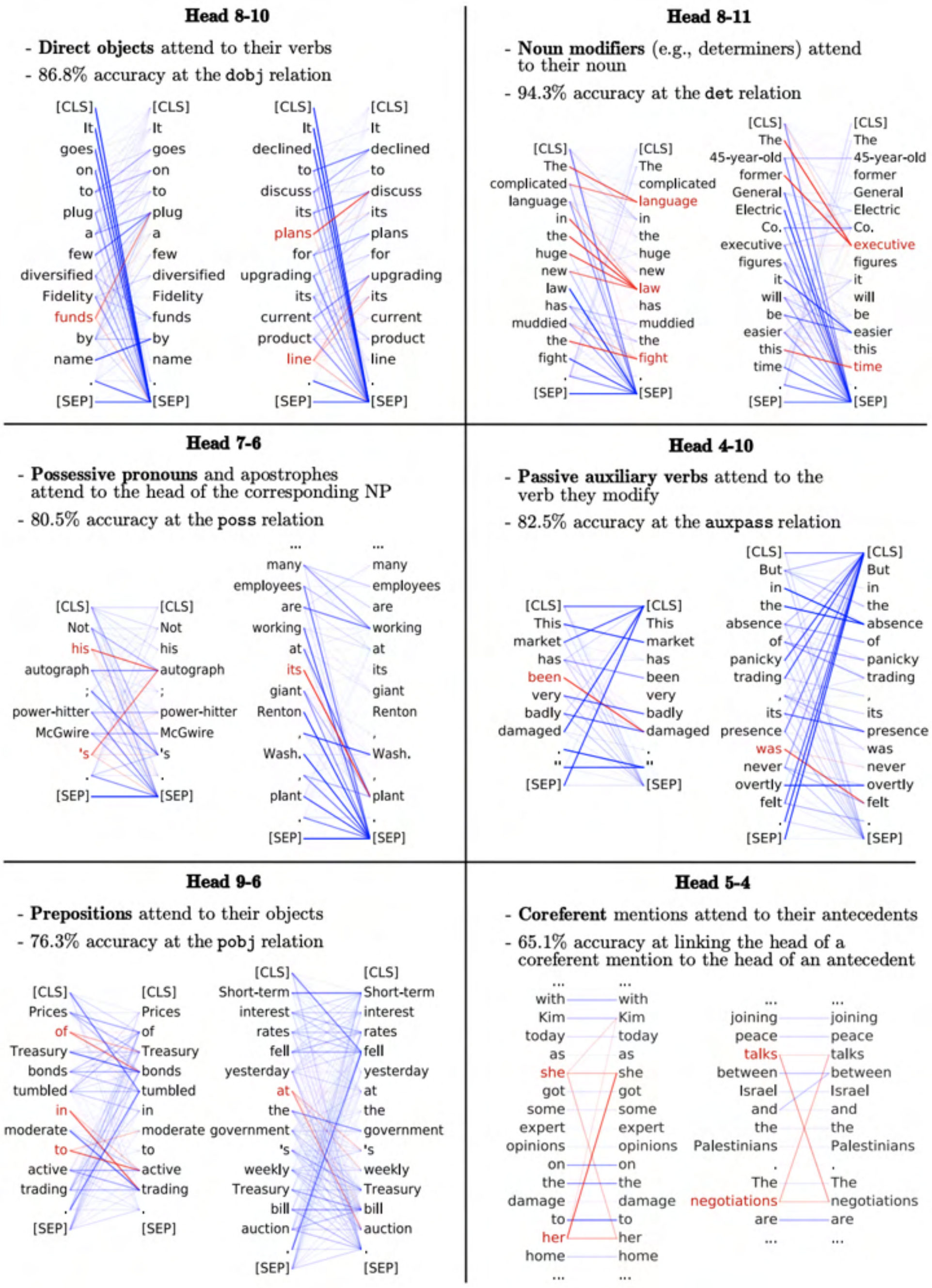

Head 8-10
- Direct objects attend to their verbs
- 86.8% accuracy at the dobj relation
[CLS] It goes on to plug a few diversified Fidelity funds by name . [SEP]
[CLS] It declined to discuss its plans for upgrading its current product line . [SEP]
Head 8-11
- Noun modifiers (e.g., determiners) attend to their noun
- 94.3% accuracy at the det relation
[CLS] The complicated language in the huge new law has muddied the fight . [SEP]
[CLS] The 45-year-old former General Electric Co. executive figures it will be easier this time . [SEP]
Head 7-6
- Possessive pronouns and apostrophes attend to the head of the corresponding NP
- 80.5% accuracy at the poss relation
[CLS] Not his autograph ; power-hitter McGwire 's . [SEP]
... many employees are working at its giant Renton , Wash. , plant . [SEP]
Head 4-10
- Passive auxiliary verbs attend to the verb they modify
- 82.5% accuracy at the auxpass relation
[CLS] This market has been very badly damaged . " [SEP]
[CLS] But in the absence of panicky trading , its presence was never overtly felt . [SEP]
Head 9-6
- Prepositions attend to their objects
- 76.3% accuracy at the pobj relation
[CLS] Prices of Treasury bonds tumbled in moderate to active trading . [SEP]
[CLS] Short-term interest rates fell yesterday at the government 's weekly Treasury bill auction . [SEP]
Head 5-4
- Coreferent mentions attend to their antecedents
- 65.1% accuracy at linking the head of a coreferent mention to the head of an antecedent
... with Kim today as she got some expert opinions on the damage to her home ...
... joining peace talks between Israel and the Palestinians . The negotiations are ...

Figure 6 BERT Attention Heads are specialised to different dependency relations. Figure from Clark et al. (2019: 280).

the Spectral Probe filters a sequence of embeddings into continuous frequency ‘bandwidths’.

Word-level tasks, like POS-tagging (POS) and phrase-level tasks like Dependency Parsing (Dep) are associated with different spectral filters across languages. Figure 7 shows the Spectral Profiles of mBERT embeddings in English/German/Spanish/ French/ Japanese/ Chinese in POS and Dep have different upper bounds. Transformers, across languages, across languages, appear to associate linguistic information that appears in the input at different timescales according to these linguistically-sensitive spectral bands.

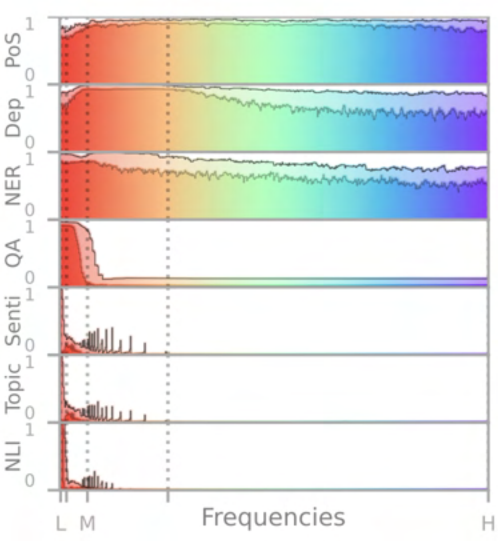


Figure 7 Spectral Filters of mBERT embeddings have different upper bounds in word/phrase-level tasks. Figure from Müller-Eberstein et al. (2022b: 7732).

The chaotic trajectories of the self-attention mechanism underpin the observed separation of different NLP tasks in Spectral Bandwidths. Word-level tasks dissociated by the lower spectral filter are lower sensitivity tasks, while phrase-level tasks are typically higher sensitivity. Hahn, Jurafsky & Futrell (2021b) introduce the notion of Block Sensitivity that measures how many disjoint subsequences can be changed individually to flip the label assigned to a phrase. Tasks at a higher sensitivity, like assigning a Universal Dependencies label, are at a higher spectral filter than lower sensitivity tasks, like identifying the relative position of a head in a parse.

Overall, Transformers share (1) a simplicity bias and (2) emergent complexity-driven curicula (reminiscent of Goldilocks effects) with MMM. As MMM is a property of the acquisitional dynamical system and Transformers simulate dynamical systems, this leads to the Sensitivity Conjecture as shown in (1).

(1) ***Sensitivity Conjecture:*** *The inductive bias of the Transformer-based LM towards simple functions and sparse variables underpins emergent syntax that approximate the behaviour of the acquisitional dynamical system.*

This preliminary synthesis of pre-existing empirical results through the theoretical lens of sensitivity as a complexity metric provides evidence for a previously unnoted convergence between MMM (as a property of the acquisitional dynamical system) and the dynamics of emergence in Transformers suggests that there are deeper formal affinities between the mechanisms developed by purely bottom-up considerations in computational linguistics and neo-emergent models of linguistic theory. This points to the productivity of postulating that neo-emergent models of grammar construction are a plausible top-down goal for improving language modelling, and motivates deeper theoretical questions about the extent to which linguistic structure can be modelled using Transformer-based architecture.

3 Research Questions

Given the convergence of research in computational and theoretical linguistics outlined in section 2.3, we can formulate two research questions to apply the theoretical predictions of MMM to Transformer LMs.

i. To what extent are Transformer Language Models sensitive to departures from Saussurean arbitrariness and contextual restrictions?

ii. How can MMM be used as a heuristic to improve the learnability of contextual restrictions in a typologically-consistent manner?

In accordance with (i), we assess in section 4 the fragility of subject-verb agreement computation in Transformer-based LMs using two creative probing strategies– causal mediation analysis and grammatical error detection. According to the two-step MMM meta-learning procedure, Transformers should be sensitive to agreement as cue for structural generalisation (it is a Departure from Saussurean Arbitrariness). Once triggered, the Transformer should recruit their inductive biases to sensitivity to learn the target contextual restriction according to a NO > ALL > SOME geometry– following a learning dynamic of overextension and retraction to reach the target distribution. We investigate the structural generalisation by measuring the extent to which agreement computation in different Transformer-based LMs can be modulated by intervening 'agreement attractors'.

To the extent that the emergent syntactic behaviour of Transformers follows the Sensitivity Conjecture, MMM is a plausible syntactico-centric top-down goal to improve the structural generalisation capabilities of Transformers in a typologically-consistent manner (section 5.1), thereby adressing (ii). In section 5.2, we begin with a theoretically-motivated discussion about the benefits of sensitivity as a syntactic complexity metric and the creative probing strategies to ascertain the structural generalisation capabilities (i.e their emergent capacity to formulate NO > ALL > SOME generalisations) in a typologically-consistent manner, as predicted by the MMM model. We utilise recent theories on neo-emergent category induction to propose a linguistically-motivated solution to help improve the structural generalisation of LMs to approximate MMM dynamics in Transformer-based LMs.

In section 5.3, in light of syntactic typological variation 'falling-out' from NO $>$ ALL $>$ SOME emergent dynamics, an MMM-perspective places a theoretical upper-bound on the effectiveness of state-of-the-art techniques for cross-lingual syntactic transfer. This motivates the dissociation of semantic and syntactic transfer learning and developing fine-grained typological evaluation datasets. I highlight how existing meta-learning techniques can be repurposed to improve structural generalisation in a typologically-consistent manner according to MMM. Augmenting the structural generalisation capacity of the Transformer in the manner may also address associated with grounding in multimodal Transformer-based architectures.

4 Case Study: Structural Generalisation in Transformer-based LMs

In this case-study, I conduct a more fine-grained analysis of the emergent subject-verb agreement capabilities of Transformer-based LMs using grammatical error detection (GED) as a diagnostic probe to assess the structural generalisation capabilities of Transformer-based LMs.[4]

4.1 Motivation: Sensitivity to Agreement Computation

Transformers recruit two MMM-based properties (Simplicity Bias and Goldilocks Effects) to exhibit the capability to distil sparse features that encode emergent syntax, like $[SG]$ and $[PL]$ number in subject-verb agreement. Following Finlayson, Mueller, Gehrmann, Shieber, Linzen & Belinkov (2021), it is possible to distil the mechanisms of agreement computation in Transformers using a technique called *causal mediation analysis*. This implicates the components of the Transformer that are involved in agreement computation by viewing each *model component* (each sentence-length embedding z in the Transformer-based LM) as a 'mediator'. Finlayson et al. measure the contribution of each embedding mediator in agreement computation by performing 'interventions' on input sentences. Finlayson et al. perform causal mediation analysis for different model sizes in Table 1.

An intervention on agreement computation is performed by switching [SG] to [PL] marking, or vice-versa, and measuring the relative change of the probability of CORRECT : INCORRECT agreement marking on the verb for each sentence-length embedding 'component' in the Transformer. The Total Effect of an intervention is measured by changing the grammatical number of the main subject (e.g author → authors) and measuring the ratio of probabilities of the originally incorrect verb form, as in (2)

[4] All code and datasets used in this section can be found in the Appendix.

Size	Layers	Embeddings Size	Headings
Distil	6	768	12
Small	12	768	12
Medium	24	1024	16
Large	36	1280	20
XL	48	1600	25

Table 1 Model Size Variants of GPT-2 used by Finlayson et al. (2021).

Given a subject prompt u and a verb v, the *total effect TE*:

(2)
$$y(u_{sg}|v) = P(\frac{\text{incorrect number}}{\text{correct number}}) < 1$$

if a Transformer-based LM predicts the correct agreement or > 1 if incorrect (Figure 8)

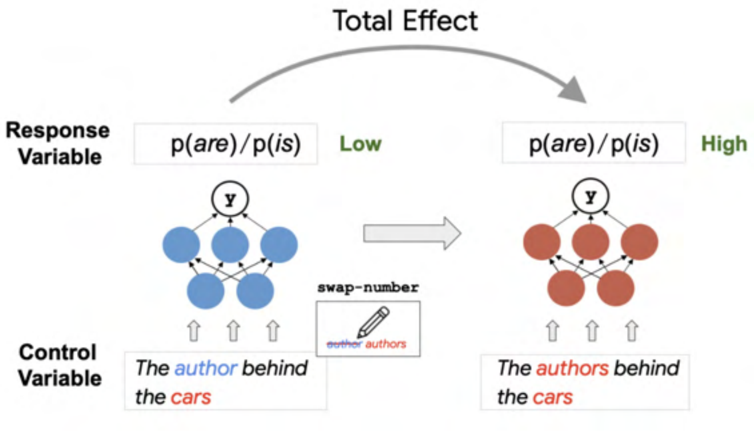


Figure 8 Total Effect of Causal Mediation Analysis. Figure from Finlayson et al. (2021: 1831).

This is measured by setting an individual embedding to the value that it would have taken if the grammatical number of a sentence was changed (although there is no swap-number intervention). This is illustrated in Figure 9.

Applying TE to the GPT-2 family, we can measure the contribution of different layers of the Transformer in agreement by measuring the *indirect total effect* (Figure 10) of swapping the number of the subject on the inflectional preferences of different model components.

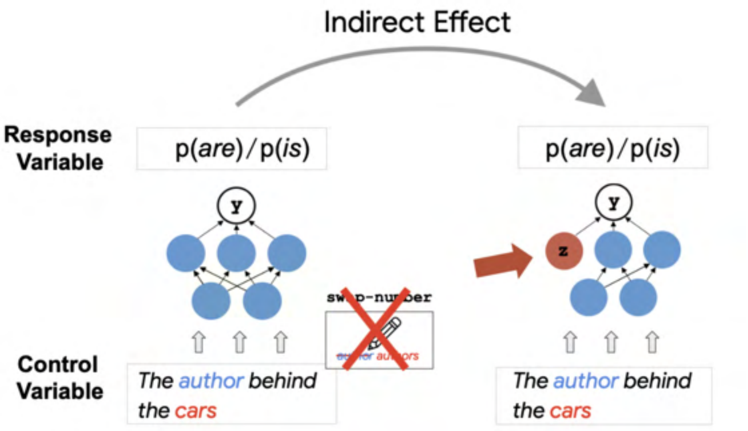


Figure 9 Indirect Total Mediation Effect. Figure from Finlayson et al. (2021: 1833).

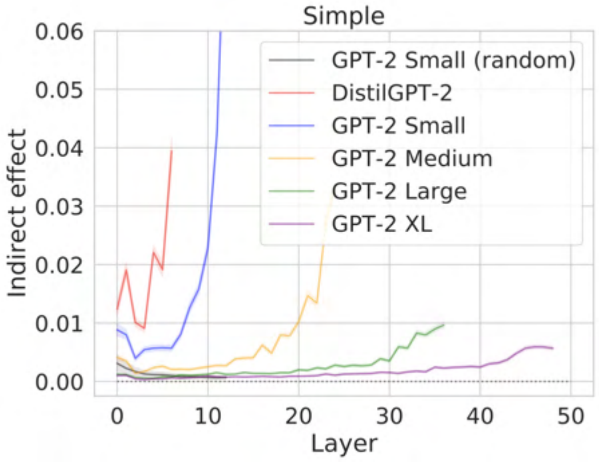


Figure 10 Indirect Total Effects across model sizes show that agreement computation is not modulated by model size. Figure from Finlayson et al. (2021: 1834).

There are two relevant results from the causal mediation analysis that motivate the probing study. First, the indirect effects suggest that differences in model size in the GPT-family do not lead to substantially different agreement mechanisms. Figure 9 shows that smaller models, like DistilGPT-2, have more syntactic knowledge concentrated in fewer embeddings that have stronger inflectional preferences in agreement computation. The figure shows the top 5% of neurons in each layer of the GPT-2 sizes in Table 1 in simple subject-verb agreement Indirect TE effects in larger models, like GPT-2 XL, are more distributed across the layers. This suggests that Transformer-based LMs exhibit sensitivity to Saussurean arbitrariness, like subject-verb agreement.

Agreement TEs are far higher than the TEs below 250 reported for gender bias in Vig, Gehrmann, Belinkov, Qian, Nevo, Singer & Shieber (2020), indicating that Transformer-based LMs show emergent syntax. Finlayson et al. (2021) find that Transformers encode morphologically-realised $[PL]$ *-s* as 'defaults' in GPT-2 and seem to be easier targets for agreement computation.

Following the Sensitivity Conjecture, Transformers do seem to recruit their MMM-like domain-general inductive biases towards learning sparse number features utilised in agreement. The causal mediation analysis suggests that emergent syntax is not modulated by model size. The insight that size does not matter for emergent syntax is significant, particularly when Large Transformer-based LMs like GPT-3 have been motivated by attempts to improve few-shot learning in downstream tasks (Brown et al. 2020). The syntactic generalisation capabilities of Transformer-based LMs seem to be at least partially dissociated from the increasing statistical correlations picked up in larger LMs and instead derive from the application of domain-general inductive biases in the self-attention mechanism.

Secondly, Finlayson et al. find that autoregressive LMs GPT-2 and Transformer XL show a divergence between local agreement (e.g. simple agreement/within a relative clause) and longer distance agreement across intervening PPs in Transformer XL Figure 11 and GPT-2 in Figure 12.

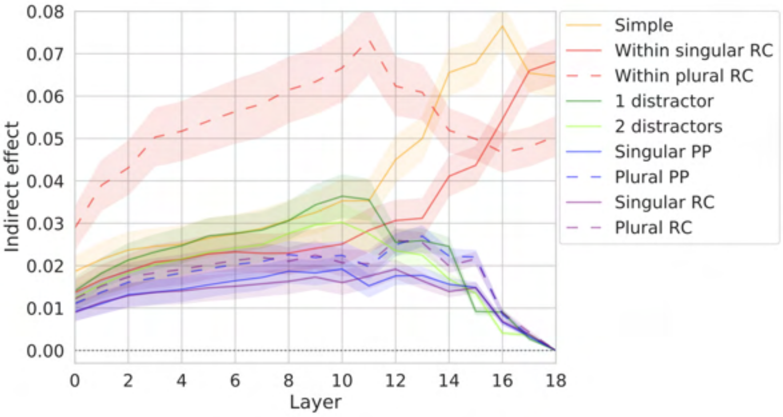


Figure 11 Transformer-XL shows a dissociation between local (shown in red/yellow) and non-local (shown in purple/green) agreement. Figure from Finlayson et al. (2021: 1834-1835).

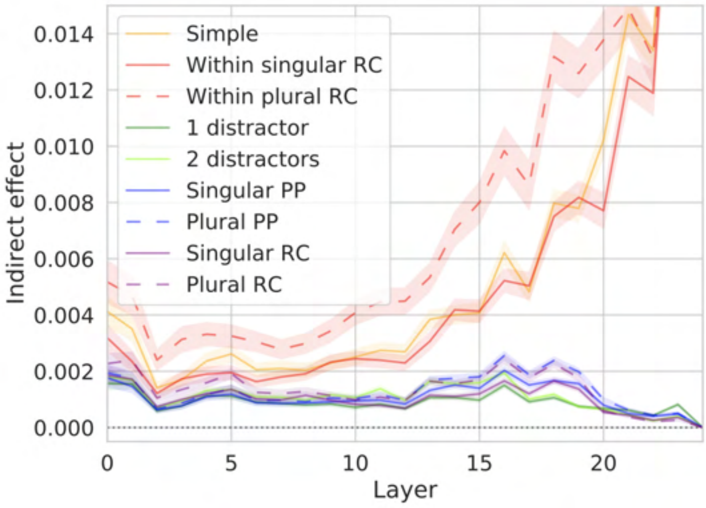


Figure 12 GPT-2 Medium shows a dissociation between local (shown in red/yellow) and non-local (shown in purple/green) agreement. Figure from Finlayson et al. (2021: 1834-1835).

This bipartite behaviour is not observed in XLNet in Figure 13, where the indirect effect contour is similar across local and non-local agreement. All three figures show the natural indirect effects of the top 5% of sentence embeddings in each layer.

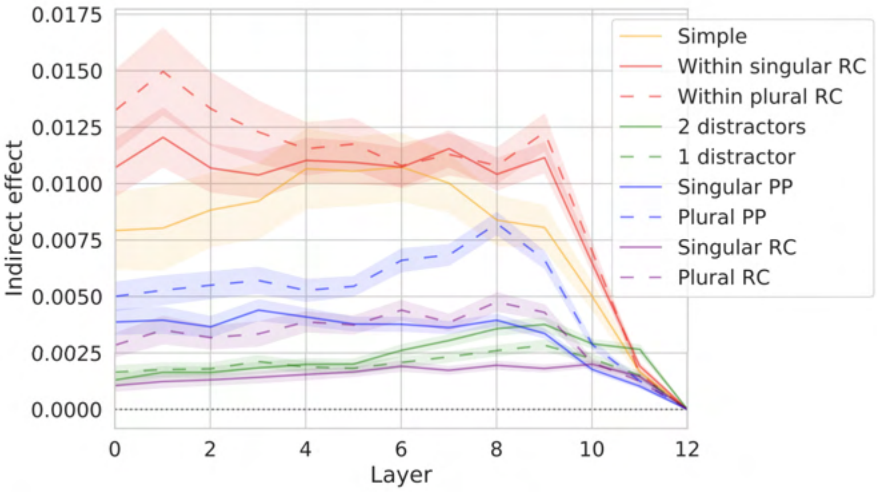


Figure 13 XLNet does not show a dissociation between local (shown in red/yellow) and non-local agreement (shown in purple/green). Figure from Finlayson et al. (2021: 1835).

The causal mediation analysis indicates that Transformer agreement computation is modulated by interfering attractors. We explore this in section 4.2.

4.2 *Grammatical Error Detection (GED) Probing*

I analyse the results of the GED probing experiment conducted by Davis, Bryant, Caines, Rei & Buttery (2022), who shared (see Appendix) their per-layer probing results across the agreement attraction conditions for five Transformer-based LMs: BERT, RoBERTa, ELECTRA, GPT-2 and XLNet. GED is a probing technique that measures the extent that Transformer-based LMs can detect ungrammatical tokens.

Contextual embeddings are extracted for every token in a sentence. All models have 12 layers. The GED probe is a token-labelling probe trained per-layer to classify the token-level embedding as grammatical or ungrammatical. This differs from previous TSE methods in two crucial respects. It does not assume errors are located on certain tokens in a sentence, allowing us to assess where the probe believes a grammatical error is located in a sentence. Error detection is assessed using contextualised embeddings, instead of targeting [MASK] tokens in BERT or RoBERTa or using minimal pair sentence-level datasets. High probing performance is indicative that the pre-trained model has encoded the relevant features in the contextual representations of VERB, which implicitly includes the subtasks of classifying contextual embeddings as verbs and subject NPs and learning the [SG] and [PL] number features.

The GED probe shows that the emergent subject-verb agreement generalisations are greatly perturbed by Agreement Attraction effects between the verb and its target subject.

4.2.1 *Methodology*

The GED probe is trained on a subset of the L2 English learner W&I-FCE dataset, comprised of Write & Improve + LOCNESS (W&I) corpus (Bryant, Felice, Andersen & Briscoe 2019) and the First Certificate in English corpus (FCE) (Yannakoudakis, Briscoe & Medlock 2011). Using the ERRANT annotation framework – a standardised grammatical error annotation toolkit that extracts edits from learner errors and corrections (Bryant, Felice & Briscoe 2017), Davis et al. restrict training data to corrected subject-verb agreement errors (tagged `R:VERB:SVA` in the ERRANT framework) in the W&I-FCE dataset, leaving 1936 sentences for training and 142 sentences for validation.

Probes are evaluated using a technique called *targeted syntactic evaluation (TSE)*. TSE methods, like CLAMS (Mueller, Nicolai, Petrou-Zeniou, Talmina & Linzen 2020) and BLIMP (Warstadt, Parrish, Liu, Mohananey, Peng, Wang & Bowman 2020), evaluate LMs on minimal pairs of grammatical and ungrammatical sentences to evaluate whether LMs can detect specific grammatical contrasts. In this case, the TSE stimuli for agreement, developed by Marvin & Linzen (2018), measure the relative probability of the matrix verb (the critical region of subject-verb agreement) agreeing with the grammatical number of NP_{subject} across agreement attractor conditions with intervening subject/object relative RCs, PPs or in coordination. The full set of stimuli is described in Table 2 below.

Structure Description	Example
Simple Agreement	John **laughs/*laughs**
In a sentential complement	John knows the dog **play/*plays**
Across a PP	The vase in the gallery **breaks/*break**
Across a subject RC	The cat that chases the mouse **runs/*run**
In a short VP coordination	John smiles and **laughs/*laugh**
Across an object RC	The mouse that the cats chase **runs/*run**
Within an object RC	The cat that the mouse **eats/*eat**
Across an object RC (null complementiser)	The mouse the cats chase **runs/*run**
Within an object RC (null complementiser)	The cat the mouse **eats/*eat**

Table 2 Table of Agreement Attraction Minimal Pair Probing Stimuli (Marvin & Linzen 2018).

Davis et al. (2022: 364) compare probes to a VERB-ONLY baseline which incorrectly tags all verbs as ungrammatical, with an average baseline score of 0.43. F1 scores on the Marvin & Linzen (2018) stimuli are reported, which equally measures the Transformer-based LM's precision in classifying token-level embeddings (correctly/incorrectly as false positives) as ungrammatical and its ability to correctly recall the token-level embeddings that were classified in the W&I-FCE dataset as ungrammatical (measuring false negatives).

4.2.2 Results

As illustrated in Figure 14, ELECTRA (orange) achieves a near-perfect F1 score across all agreement attractor conditions. GPT-2 (green) is below baseline on all metrics. BERT-probes perform well for most constructions (although performance is slightly lower in object RCs), especially in layer 12, suggesting that the token representation from BERT models already encodes a lot of information related to SVA before any further fine-tuning. BERT's encoded syntactic information about POS tagging is known to be 'catastrophically' forgotten during finetuning, while syntactic information related to dependency and constituency parsing is improved. We discuss this further in section 5.2.

F1 scores for RoBERTa (red) drop for sentences with subject-relative clauses, prepositional phrases, and object-relative clauses (agreement within the clause). Performance is much worse in sentential complements, subject relatives, and short/long VP-coordination. In XLNet (purple/brown), F1 scores are around baseline for PP and basic subject-verb agreement; much worse for intervening RC (within RC, and with no complementiser); and higher than the baseline for across RCs (with/without complementiser). Sentences with attractors in these models compromise meaning independence when processing the agreement relation.

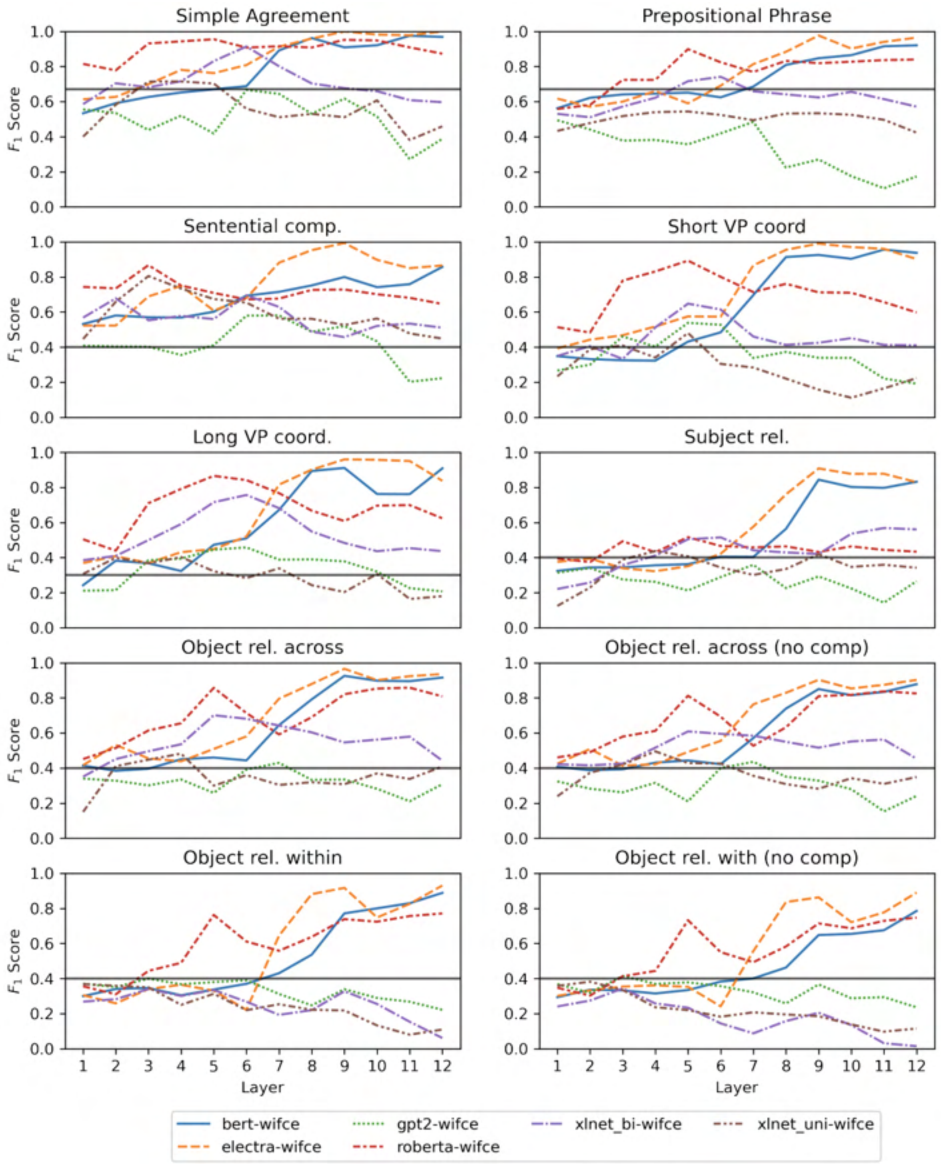


Figure 14 Results of the Grammatical Error Detection Probe (Davis et al. 2022), trained on the W&I-FCE dataset, on the agreement attraction targeted syntactic evaluation conditions (Marvin & Linzen 2018). Data courtesy of Chris Davis and Andrew Caines.

Psycholinguistic evidence suggests that agreement attractors modulate agreement computation for only 13% of complex NPs (Acuña-Fariña 2012: i.a.). Marvin & Linzen (2018: 1197-1198) conduct an acceptability judgement test on Amazon Mechanical Turk, giving human-participants minimal pairs of the agreement attractor set which only dropped to $82 - 88\%$ accuracy in non-local agreement (e.g. across RCs and PPs). These low-probability Agreement Attraction effects have been argued to be the product of erroneous grammatical computation in agreement systems (Franck, Sadri Mirdamadi & Kahnemuyipour 2020, Bhatia & Dillon 2022).

While it was not possible to directly compare the F1 scores of the model with available human baselines, structural generalisations in RoBERTa and XLNet are modulated to a far greater extent by agreement attraction effects, particularly in relative clause and coordinative environments. This reduced performance in long-distance contexts differs from human grammatical competence

4.2.3 *Analysis: Limited Structural Generalisation*

I highlight four factors that potentially modulate the sensitivity of Transformer-based LMs to agreement-attraction effects:

i. *Frequency:* While BERT can generalise well to SVA pairs that never occur in training (which indicates a degree of rule-governed behaviour), verb frequency and the frequency of alternative inflections in training data are causally implicated in the predictions BERT makes at inference time– these strong training priors make it harder for BERT to estimate the grammatical number of infrequent lexical items (Wei, Garrette, Linzen & Pavlick 2021).

ii. *Spurious Correlations:* Transformer models exploit increasingly smaller spurious statistical correlations in the dataset– and common strategies like balancing datasets may be bound to 'throw the baby out with the bathwater' (Schwartz & Stanovsky 2022). Transformers lack relevant domain-general knowledge to distinguish these spurious correlations from genuine causation (Eisenstein 2022).

iii. *Effect of Long Distance:* Transformers exhibit a brittle performance in long-range embedded dependencies (Lakretz, Desbordes, Hupkes & Dehaene 2022).

iv. The *compositional generalisation* abilities of Transformers are highly sensitive to the characteristics of the training data. Making simple changes to training data distribution (e.g. adding more varied primitives in the data) allows Transformers to generalise compositionally (Patel, Bhattamishra, Blunsom & Goyal 2022).

4.2.4 *Benefits of MLM*

The GED probing results show that autoregressive LMs, like GPT-2 and XLNet, perform only on-par with the VERB-only baseline and do not encode robust information for SVA error detection. The overall F1 scores for autoregressive LMs (GPT-2, XLNet), MLM models (BERT, RoBERTA) and ELECTRA across all agreement attractors are shown in Figure 15. The MLM pre-training objective in BERT and RoBERTa encodes inductive biases that are relevant to the detection of SVA errors, compared to the autoregressive language modelling objective in GPT-2 and XLNet. Both BERT and ELECTRA encode information related to SVA errors in the middle-to-late layers, while ROBERTA seems to encode information earlier in the model, while ELECTRA's Replaced Token Detection Objective is more aligned with GED, which potentially explains its greater sensitivity to encode syntactically discriminative information.

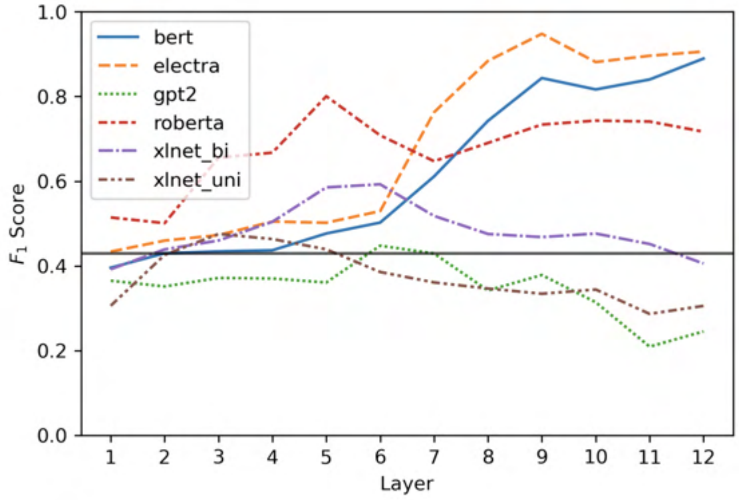
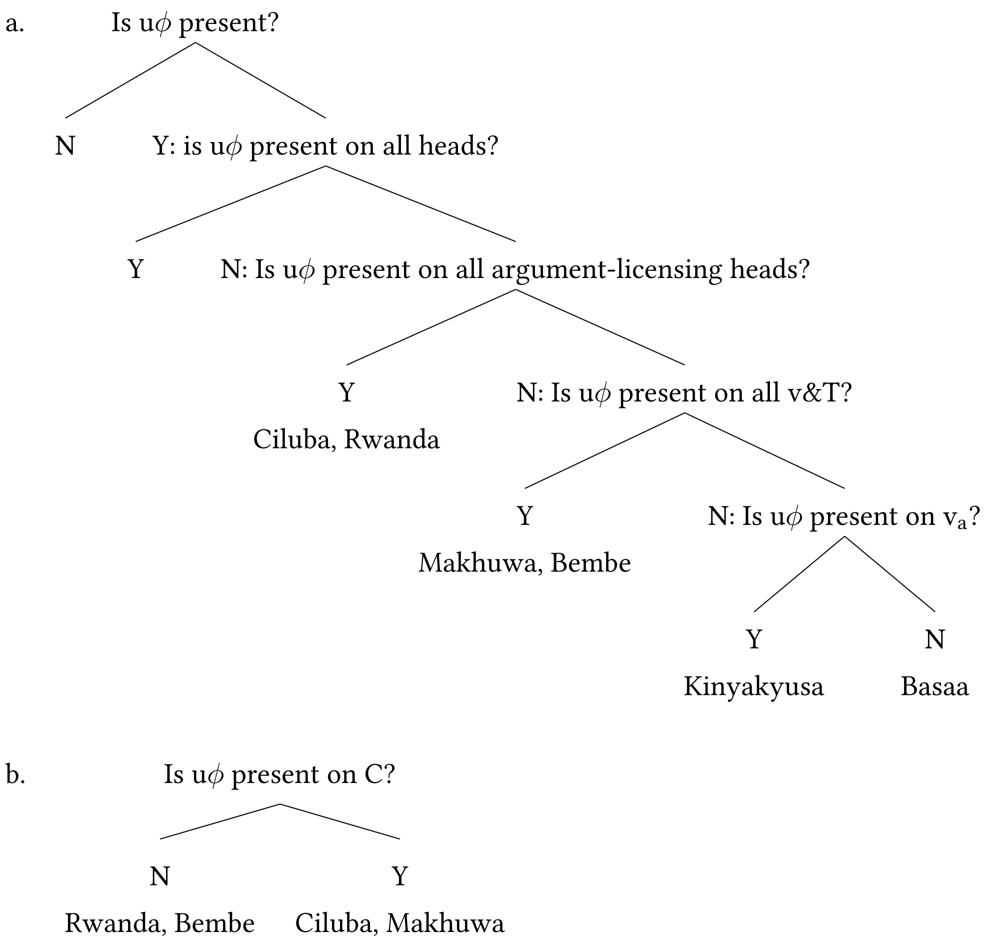


Figure 15 Masked Language Models and ELECTRA perform far better than autoregressive LMs. Figure based on data shared by Andrew Caines and Chris Davis.

MLM objectives are biased towards extracting both statistical and syntactic dependencies using random masks. While uniform [MASK] lacks task-specific supervision, the MLM objective can recover latent variables. The ability for MLM log-probabilities to recover dependencies between these variables implicitly supports emergent syntax: Transformers with an MLM pre-training objective can outperform classical unsupervised parsing methods when a minimum spanning tree is formed on the implied statistical dependencies produced by the attention mechanism (Zhang & Hashimoto 2021).

Empirical evidence of the fragile computation of agreement demonstrates the poor domain-general structural generalisation capabilities of Transformer-based Language Models. Once a human learner postulates (uninterpretable) ϕ-features, $[u\phi]$ (triggered by 'departures of Saussurean arbitrariness') the learner restricts the domain of $[u\phi]$ on syntactic heads– initially extending it to all syntactic heads in the target domain before successively restricting the target domain based on further input. These acquisitional dynamics yield the Parameter Hierarchy for Bantu agreement in (2) (van der Wal 2022: 242) – we return to Parameter Hierarchies in section 5.3.

(3) Emergent Parameter Hierarchy for agreement in Bantu languages. The NO $>$ ALL $>$ SOME contextual restriction of (uninterpretable) ϕ-features, uϕ, on specified syntactic effects leads to robust agreement computation. Feature uninterpretability is assumed to trigger agreement computation in the Minimalist Program (Chomsky 2000), but is not relevant for our purposes. Figure from van der Wal (2022: 242).


a.
Is uϕ present?
N
Y: is uϕ present on all heads?
Y
N: Is uϕ present on all argument-licensing heads?
Y
Ciluba, Rwanda
N: Is uϕ present on all v&T?
Y
Makhuwa, Bembe
N: Is uϕ present on v_a?
Y
Kinyakyusa
N
Basaa
b.
Is uϕ present on C?
N
Rwanda, Bembe
Y
Ciluba, Makhuwa

In MMM terms, the absence of the ability to perform NO $>$ ALL $>$ SOME learning dynamics means that Transformers are unable to replicate the contextual-restriction of $[u\phi]$ to its target domain, leading to an overgeneration of agreement attraction effects modulated by an over-sensitivity to spurious correlations. Emergent syntax must be distilled more efficiently to improve the structural generalisation capabilities of Transformer-based LMs. In section 5, I propose techniques that can do so with reductions in corpus size for pre-training and benefits in low-resource settings.

5 Discussion: MMM as a heuristic for Language Modelling

5.1 Top-Down Goals in Language Modelling

Striking a correspondence between the pre-training of Transformer language models and the dynamics of grammar construction, according to the predictions of MMM and DST, allows us to interpret empirical probing results by linking MMM-driven acquisition pathways to the inductive biases of the neural architecture. Although there may be confounding variables that impinge on the results of the GED probing and causal mediation analyses of agreement, they do indicate that Transformer-based LMs exhibit domain-general inductive biases that are recruited to replicate properties attributed to the acquisitional dynamical system of the human learner, such as sensitivity to departures from Saussurean arbitrariness and Goldilocks Effects.

It is conceptually and theoretically desirable to translate the predictions of MMM and DST into top-down language modelling goals, which guide the process of LM along a radical cline of increasing linguistic sensitivity, taking LMs to a position of having the competence of modelling possible and probable languages. MMM is a more abstract framework couched in the background assumptions of generative grammar. This means that it is not possible to fully replicate the MMM model in Transformers in the absence of an intensional characterisation of formal features in contemporary syntactic theory[5]. Instead, MMM is a *heuristic* for computational linguists to discriminate the most salient measures of syntactic complexity; probing strategies; and typological datasets. As a universal meta-learning strategy that underpins emergence in natural language, it allows us to identify a constrained set of pre-training and meta-learning techniques to augment the basic Transformer LM to approximate the NO > ALL > SOME human emergent structural generalisation capabilities across languages.

As the technological tools for the injecting syntactic behaviour of MMM and DST into Transformer-based LMs are in their infancy, I argue MMM can be used as top-down heuristic to inform seven practical improvements in NLP practice that should yield better syntactic generalisation (section 5.2) in a typologically-equitable manner (section 5.3). We extend the methodology of Collier, Liu & Shareghi (2022), which triangulates between philosophical ontology, cognitive psychology and computational linguistics to develop cognitively-motivated evaluation metrics to close the gap with normative human performance on symbol grounding (Harnad 1990) and common-sense reasoning. This top-down approach is adjunct to intrinsic evaluation metrics, like perplexity. While these need to be reconciled with bottom-up model performance-driven considerations, this approach presents a narrow pathway for future linguistically-motivated LM engineering to allow Transformers to fully capitalise on their inductive biases for emergent syntax.

[5] Matt Tyler and Theresa Biberauer (p.c.)

5.2 Improving Structural Generalisation

5.2.1 Measures of Syntactic Complexity: Surprisal, Chomsky Hierarchy and Sensitivity

Sensitivity is computable and theoretically-motivated complexity measure that can relate the inductive biases of the Transformer to its emergent syntactic properties. It can tractably estimate the relative complexity of NLP tasks, unlike Kolmogorov complexity which is well-defined and computable only in the asymptotic limit. Functions with low sensitivity also have a low Kolmogorov-complexity– this is a metric applied in neo-emergent approaches to compare the 'depth' of emergent parameter hierarchies (Biberauer, Holmberg, Roberts & Sheehan 2014b). I now argue that sensitivity is a more suitable complexity metric for emergent syntax than the Chomsky Hierarchy and information-theoretic intrinsic evaluation metrics.

The Chomsky Hierarchy is a containment hierarchy of formal languages (Chomsky 1956), comprised of regular languages that generate left/right-linear production rules $A \rightarrow a$ and $A \rightarrow aB$ which can traditionally model (morpho)phonological rules (Chomsky & Halle 1968); context-free languages which can additionally generate binary-branching phrase-structure rules, $A \rightarrow BC$; and 'mildly' context-sensitive languages– which can characterise certain computational processes (Joshi 1985) as summarised in Figure 16. As consistent with early computational morphology modelled using finite-state transducers (Kaplan & Kay 1994: i.a.), most morphophonological processes are modelled in the 'sub-regular' region (Chandlee, Eyraud & Heinz 2014, Heinz 2018, Danis & Jardine 2019).

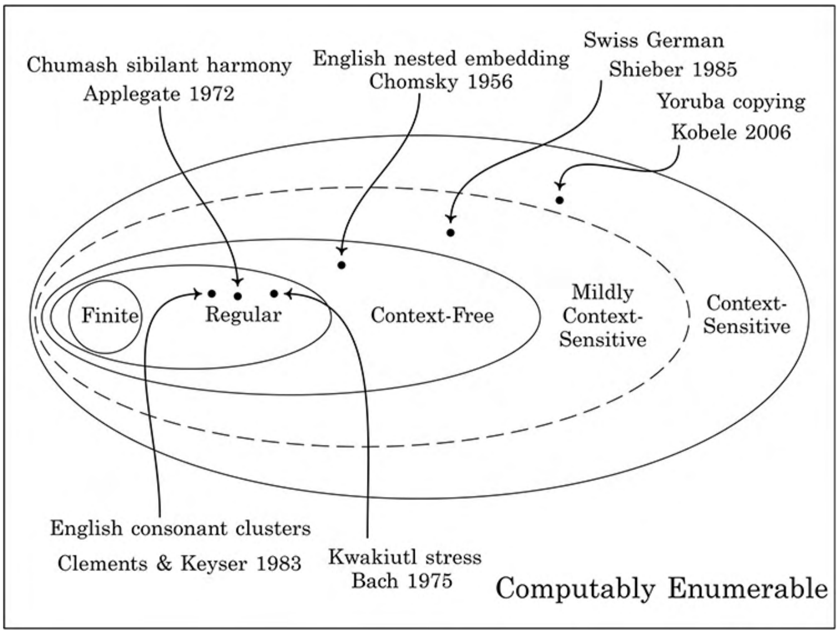


Figure 16 Location of Grammatical Processes on the Chomsky Hierarchy of Formal Languages. Figure from Rawski & Heinz (2019).

While the standard linguistic interpretation of the Chomsky Hierarchy in Figure 16 is consistent with early pre-*Aspects* Generativism, the class of formal languages described by a family of parsing techniques called Minimalist Grammars that are the most faithful to current Chomskyan commitments is exactly the class described by multiple-context-free grammars (Stabler 2011: 624):

$$\begin{gathered}\text{Context-Free} \subset \text{TAG} \equiv \text{Categorial Grammar (CCG)} \\ \subset \text{Multiple CFGs} \equiv \text{MG} \subset \text{Context-Sensitive}\end{gathered} \tag{4}$$

Investigations of the computational complexity of Transformers, which impose formal bounds on the expressivity of the Transformer encoder in terms of the Chomsky Hierarchy, have found that it is (paradoxically) limited in recognising unbounded hierarchical languages and even certain regular formal languages despite its improved surface performance. This is exemplified in Table 3.

Formal Language	Circuit Complexity	Y/N	Description	Reference
PARITY	Non-AC^0	N	The language of bit strings with an odd number of 1s	Hahn (2020)
2Dyck	Non-AC^0	N	The language of ‘correctly bracketed words’ consisting of two types of brackets (‘(’,‘)’ and ‘[’,‘]’).	Hahn (2020)
Dyck-1 Majority	AC^0 Boolean Circuits of Constant Depth	(Y) by self-attention variant **generalised unique hard attention**	DYCK-1 = {‘(’,‘)’} MAJORITY = set of strings with at least as many 1s and 0s	Merril et al (2022) Hao et al (2022)
Counter Languages	N/A	Y	Non-Deterministic FSA with additional memory to hold a non-negative integer that can be incremented/decremented	Bhattamishra et al (2022)
$\text{Dyck}_{k,D}$	N/A	Y	Subset of Dyck_k depth bounded by D This is argued to capture the (finite) bounded hierarchichal structure of natural language	Yao et al (2021)
Boundary: The upper bound for fixed-precision Transformer Encoders is first-order logic with counting qualifiers foc[+;MOD] Chiang et al (2023)				

Table 3 Transformers can recognise the proper subset of regular and subset of deterministic context-free languages (counter automata). The upper-bound of the expressivity of Transformer encoders with fixed-precision can be modelled by characterising strings using logical formulae (Chiang et al. 2023).

There are four reasons why sensitivity should be a preferred metric of emergent syntax over the Chomsky Hierarchy:

i. The formal languages used to bound the complexity of the self-attention mechanism are more naturally suited to recurrent architectures– as finite-state automata are formal abstractions of RNNs (Bhattamishra et al. 2022: 11). This view of computational complexity does not isolate the inductive biases of Transformers that are recruited in language-specific tasks.

ii. While the Chomsky Hierarchy has been cited to be a useful formalism to characterise locality in Generative Linguistics (Avcu & Rhodes 2022), it may be too empirically restrictive (Chaves & Putnam 2022).

iii. The Chomsky Hierarchy does not measure the difficulty for a Transformer to achieve high accuracy on tasks given a realistic distribution of output– as it is not defined for individual inputs (Hahn et al. 2021b: 891).

iv. There is evidence to suggest that 'shallow' Transformers can learn 'shortcuts' to a wide variety of automata structures during pre-training (Liu, Ash, Goel, Krishnamurthy & Zhang 2022).

This 'shortcut behaviour' of Transformers in learning formal languages is expected under the Sensitivity Conjecture as a 'minimal means' used by the Self-Attention mechanism to pay attention to salient aspects of the input. Bhattamishra et al. (2022: 21-22) establishes that sensitivity is a capacity measure that derives bounds on the structural generalisation capabilities of Transformers. Transformer-based LMs are more likely to overfit and achieve poorer generalisation performance in higher-sensitivity functions. Conversely, functions with lower sensitivity can be learnt with better sample efficiency. Sensitivity is a capacity bound for the upper bound structural generalisation in Transformers: high sensitivity functions have high entropy, but the converse is not necessarily true (Bhattamishra et al. 2022: 21-22). The Sensitivity Conjecture characterises the fundamental patterns of emergence that underpin the poor structural generalisation of Transformer-based LMs.

The standard approach for interpreting syntactic emergence relies on the information-theoretic notion of surprisal. Surprisal is closely related to intrinsic evaluation metric perplexity, which determines the number of guesses an LM takes to correctly predict the target word w_i:

(5)
$$\prod_{i=0}^{N} P(w_i|w_{\leq i})^{\frac{-1}{N}}$$

Surprisal is the negative log-probability $-log[P(w_i|w_1,\ldots,w_{i-1})]$ of input, $w_1, \ldots, w_n$. The central claim of this Surprisal Theory (Levy 2018, Futrell & Levy 2017, Futrell, Wilcox, Morita, Qian, Ballesteros & Levy 2019, Wilcox, Levy, Morita & Futrell 2018, Thrush, Wilcox & Levy 2020: i.a.) of emergent syntax is that the human-likeness of model predictions can be attributed to surprisal differences of

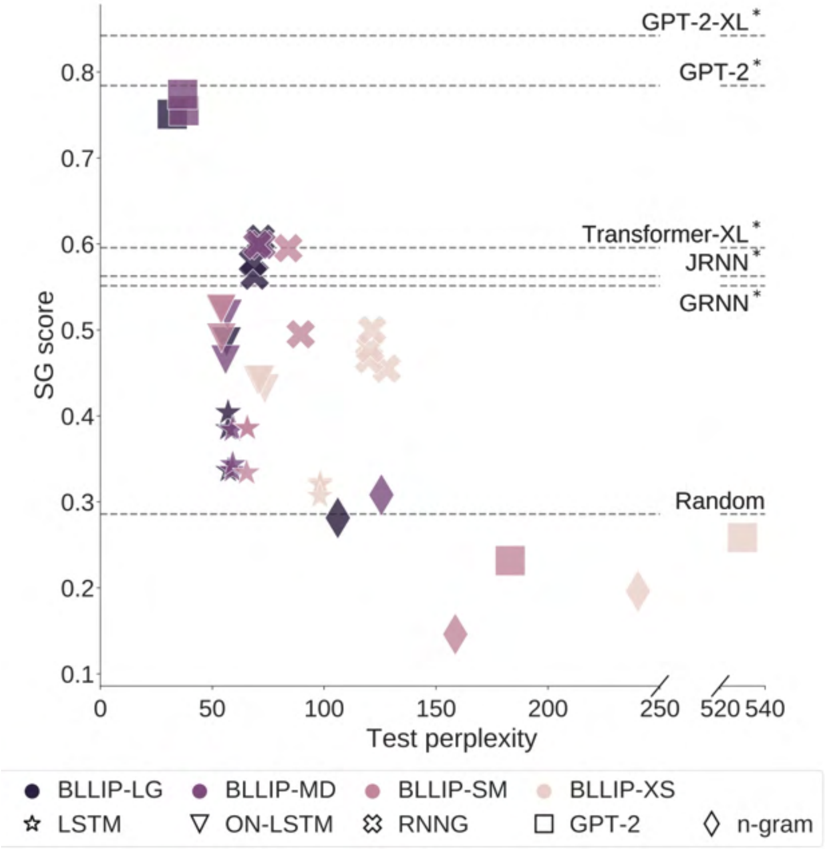


Figure 17 Structural Generalisation is partially dissociated from Perplexity. Figure from Hu et al. (2020).

minimal pairs of syntactic constructions. Yet, there are three limitations of Surprisal Theory.

First, Syntactic Generalisation performance can be dissociated from standard information-theoretic metrics of LMs, like perplexity (the inverse of surprisal). Substantial differences in generalisation depend on pre-training objectives of the Transformer-based model (Hu, Gauthier, Qian, Wilcox & Levy 2020), which cannot be modelled using surprisal. Additionally, Surprisal Theory does not assess the empirical effectiveness of different entropy estimators for syntactic distributions: Arora, Meister & Cotterell (2022) indicate that the use of poor entropy estimators can lead to an over-estimation of effect sizes in information-theoretic studies.

Moreover, Surprisal is not sufficient to account for attested contiguity effects in Language. While Hahn, Degen & Futrell (2021a) suggest information locality is a surprisal-derived functional constraint that drives systems towards harmonic OV-VO word order, the MMM dynamics in human acquisition yield more abstract contiguity effects which cannot be characterised in terms of optimal coding from an information-theoretic perspective. *The Final-over-Final Condition (FOFC)* (Biberauer, Holmberg & Roberts 2014a, Sheehan, Biberauer, Roberts & Holmberg 2017) is typological contiguity restriction which requires structurally adjacent syntactic heads to bear a diacritic $\wedge$ to prohibit the derivation of typologically unattested word orders like $*$VOAux (see Figure 18). Contiguity is realised as a reflex of MMM to create contiguous syntactic domains that share a formal feature with no intervening 'on/off' patterns.

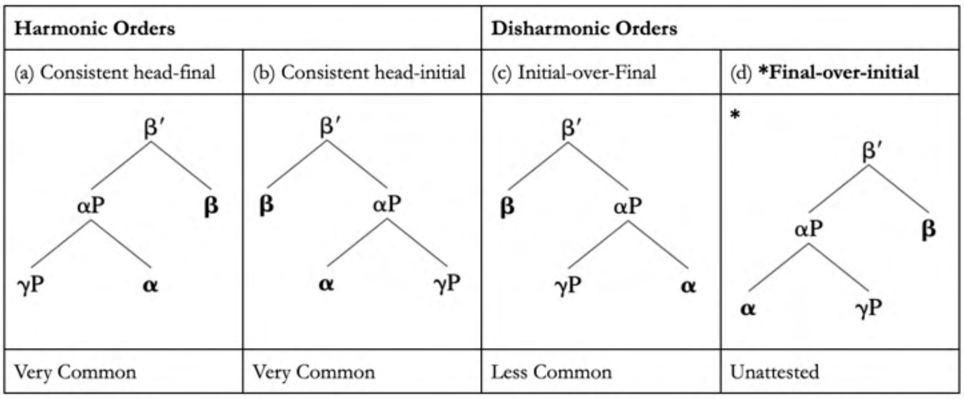


Figure 18 The Final-over-Final-Condition (Biberauer et al. 2014a, Sheehan et al. 2017) is an abstract contiguity restriction that rules out typologically-unattested *final-over-initial word disharmonic orders. FOFC cannot be accounted for by surprisal.

Typologically Unattested Word Order	FOFC-Incompliant Structure
*VOAux	*$[_{AuxP}[_{VP}$ V DP] Aux]
*VOC	*$[_{CP}[_{TP}$ T VP] C] or * $[_{CP}[_{TP}[_{VP}$ V O] T] C]
*C-TP-V	*$[_{VP}$ $[_{VP}$ C TP] V]
*Det/N PP P	*$[_{PP}[_{NP/DP/}$ D/N PP] P]
*Num-NP-Dem	*$[_{DemP}[_{NumP}$ Num NP]DEM]
*Pol-TP-C	*$[_{CP}[_{PolP}$ Pol TP]C]

Table 4 Examples of FOFC Violations.

Finally, Wilcox, Futrell & Levy (2022: 35-37) claim that emergent syntax is driven by 'data-likelihood', as there is no 'obviously strong natural language syntax-oriented inductive bias' in the Transformer. Dependencies between token sequences are argued to underpin Transformer's emergent capabilities to learn locality constraints because of a high data-likelihood to shift probability mass towards the target distribution in the training data. This account of emergent syntax is unrestrictive and unpredictive: it does not explain how Transformers recruit their inductive biases to formulate generalisations given the poverty-of-stimulus in the training data. Surprisal Theory does not account for which facets of the input the Transformer attends to, nor the order in which the emergence of syntactic information follows. The Sensitivity Conjecture is a more fundamental measure of syntactic emergence than Surprisal – which can only mirror functional information locality effects in processing that is partially dissociated from structural generalisation, rather than elucidating deeper learnability parallels with linguistic competence.

5.2.2 Probing Techniques for Emergent Syntax

The effectiveness of probing for emergent syntax has, recently, been disputed. While probing can clarify the opacity of Transformer representations, Maudslay, Valvoda, Pimentel, Williams & Cotterell (2020) argue that there should be no inherent difference between probe design and the model designed for the task in computational linguistics– after finding that a simple parser with a lightweight parameterisation was able to identify more syntax in BERT than the Structural Probe (Hewitt & Manning 2019). Classical probes may memorise information from the dataset without evaluating representations in the Transformer (Belinkov 2022).

Immer, Torroba Hennigen, Fortuin & Cotterell (2022) present a Bayesian framework to ensure that probes quantify the inductive bias towards linguistic differences between embeddings by selecting probes that are the right complexity. Spectral Probing (Müller-Eberstein et al. 2022b) allows us to gain insights into how Transformers recruit their inductive biases according to emergent curricula, enabling quantitative comparisons between the linguistic intuitions underlying tasks. However, this approach is limited to high-resource languages, where datasets for all tasks are available.

Viewing Transformers as dynamical systems allows us to relate the creative probing strategy, Grammatical Error Detection, and Causal Mediation Analysis used in section 4 directly to the inductive biases of the Transformer. GED extracts linguistic information from the Transformer by harnessing ungrammaticality detection as a window into the linguistic competence of the model. GED does not learn additional parameters as measures the effect of perturbations to the input sentence and it is compared to a random baseline. Causal Mediation Analysis was introduced as an alternative to classical probing strategies (Vig et al. 2020): performing interventions to the input sentences to measure the changes to a set of continuations elucidates how model size does not affect significantly increase syntactic performance that directly implicates model *causation* (not correlation) and does not introduce additional confounds in a paradigm that is reminiscent of minimal pair acceptability tests used in psycholinguistics.

Future extensions of Causal Mediation and GED, which has a close relationship to human grammaticality judgements, can potentially harness Sensitivity as a complexity measure to develop strategies that directly quantify how the MMM-reminiscent inductive biases underpin emergent syntax.

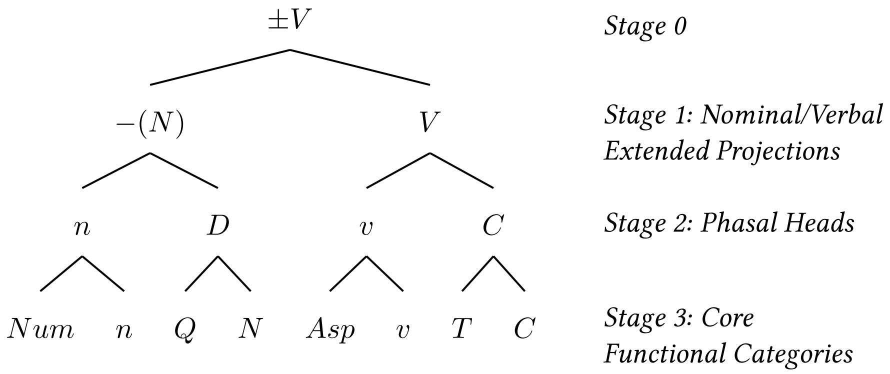


Figure 19 The Sprouting of the Syntactic Category System (Bosch 2023).

5.2.3 *Replicating Category Sprouting: Augmenting Complexity-Driven Curricula*

Transformer-based LMs need to improve the robustness of emergent category induction to improve syntactic generalisation to avoid erroneous influence by agreement attractors.

A recent neo-emergent alternative of maturation in grammar construction called ‘Sprouting’ proposes that the representational *syntactic category system (SCS)* is comprised of a universal developmental shift from an initial state with a unified category $\mathscr{C}$ that successively divides into nominal and verbal extended projections, $\mathscr{C}_{EP} = \{\mathscr{C}_{\text{nominal}}, \mathscr{C}_{\text{verbal}}\}$, then *‘phasal’* heads $\mathscr{C}_{\text{Phasal}} = \{C, v, ...\}$ – which are involved in cyclic symbolic-computation (Chomsky 2001) –, the ‘core functional categories’ (CFCs) and finally more fine-grained projections based on the input (Bosch 2023). This neo-emergent SCS is typologically-motivated by generative work on Cartography (Rizzi & Haegeman 1997: i.a.) and on the relative order of the emergence of syntactic categories in children (Friedmann, Belletti & Rizzi 2021: i.a.).

Sprouting categories abide by the successive division behaviour of MMM, as there is an assumed structural homology between the SCS and the acquisitional dynamical system. Under a top-down approach, Sprouting offers a natural solution to the empirical and theoretical issues raised in section 4.2 if we can view a discrete subset of the real vector space of the Transformer encoder states as isomorphic to a combinatorial space of discrete symbol structures.

Transformer-based LMs have not achieved state-of-the-art performance in POS tagging tasks– models do not preserve POS information which can be distilled from surrounding context (Pérez-Mayos, Ballesteros & Wanner 2021a, Pérez-Mayos, Carlini, Ballesteros & Wanner 2021b) and recent state-of-the-art performance is made by a Gaussian Hidden-Markov-Model (Zhou, Li, Li & Zhang 2022). More recently, Syntax-informed Transformer-based LMs (in Table 5) often rely on additional syntax-guided attention components to enhance the transformer– using syntax-aware attention that restricts self-attention to syntactically-relevant ‘local’ regions.

These methods require more parameters and additional syntactic parsing in downstream tasks, which severely limits the application of syntax-enhanced language

Name	Technique
SpanBERT (Joshi et al 2020)	Extends BERT by making contiguous random spans, rather than random tokens, and training the plan boundary representationsto predict the entire content of the masked span, without relying on the individual token representations within it.
Structural Scaffolding (Quian et al 2021)	Augment Transformer with 'Generative Parsing' that jointly models the incremental parse and word sequence as a part of the same sequence modeling task and a 'Structural Scaffold' that guides the language model's representation via additional structure loss that separately predicts the incremental constituency parse
Syntax-BERT (Bai et al 2021)	Generates sub-networks based on sparse masks reflecting different relationships and distances of tokens in a syntax tree, with a topical attention layer to aggregate task-oriented representations from different sub-networks
Syntax-Aware Local Attention (Li et al 2020)	Attention scopes are restrained based on the distances in the syntactic structure
Transformer Grammars (Sartran et al 2022)	Syntax LMs jointly model probability of phrase structure trees and strings of words using a STACK/COMPOSE attention mechanism

Table 5 Summary of recent Syntax-informed Transformer LMs.

models in a wide range of 'downstream' NLP tasks. Instead, the unsupervised distillation of syntactic information is possible if models are prompted to partition syntactic and contextual information (Bailly & Gábor 2020). For instance, it is possible to distil syntactic information by learning a function $f : \mathbb{R}^n \to \mathbb{R}^m$, which operates on contextualised word representations x and extracts vectors, $f(x)$, that make the structural information encoded in x more salient and discarding as much lexical information as possible (Ravfogel, Elazar, Goldberger & Goldberg 2020). The function is learnt by sampling sentences from a LM and replacing content words to obtain 'structurally-equivalent' sentences and then learning a mapping from contextualised embeddings to the words using a technique called Triplet Loss which pushes the representations of pairs from the same group closer together. Zhang, Lijie, Xiao & Wu (2022) inject syntactic structure into Transformers using a phrase-guided contrastive objective that maximises attention distributions between words in the same phrase, contrastively inducing dependency tree representations from attention distributions.

Moreover, Huebner, Sulem, Cynthia & Roth (2021) trains a scaled-down version RoBERTa on a 5 million word corpus of child data to simulate the PLD available to the child. Dubbed BabyBERTa, it receives 50 million tokens of input using a dynamic masking strategy, equivalent to the linguistic experience of a 6;0 child. When evaluated on a syntactic test suite (see Figure 20), BabyBERTa reaches comparable accuracy to RoBERTa despite having 15X fewer model parameters. Encapsulating the complexity-driven curricula sprouting dynamic, pre-training Transformers using small-scale PLD is a technique that can be extended to enable resource-efficient structural generalisation in a typologically-consistent manner.

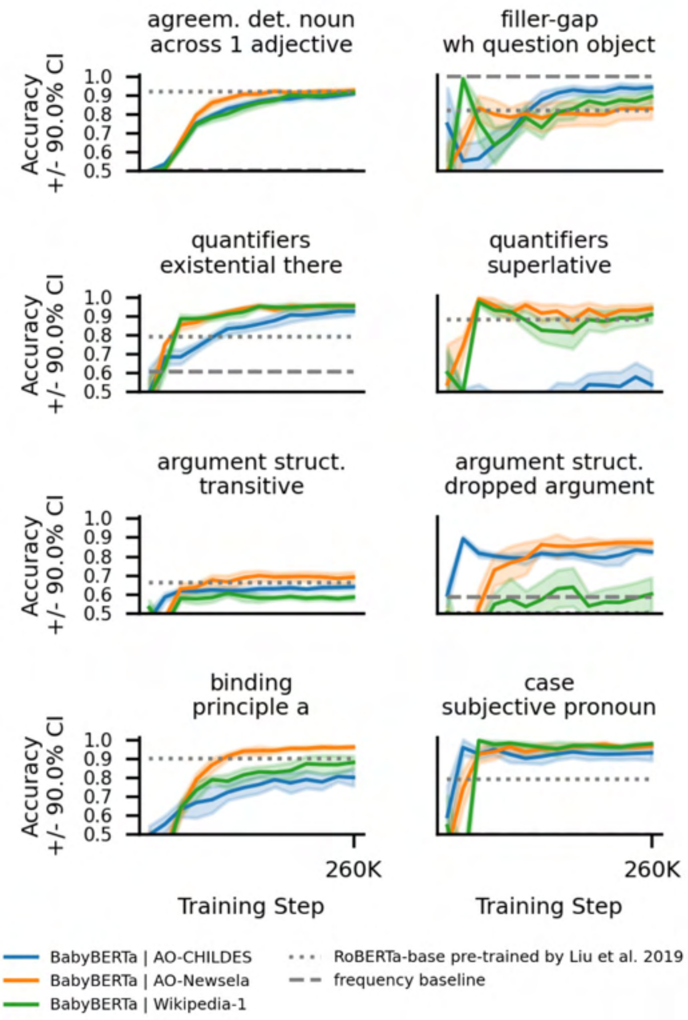


Figure 20 BabyBERTa, trained on age-ordered child corpora like CHILDES, achieves comparable accuracy to RoBERTa on a minimal pairs TSE dataset inspired by BLIMP (Warstadt et al. 2020). Figure from Huebner et al. (2021: 629).

While the precise nature of the 'mapping' between symbolic computation and the self-attention mechanism is unclear from a top-down perspective, these two families of techniques that from a 'bottom-up' perspective are potentially relevant. Guided by MMM-considerations, I hope to augment Huebner-style pre-training with some unsupervised syntactic distillation mechanism to support robust syntactic generalisation would allow the model to exhibit 'Sprouting' emergent syntactic categories according to a universal syntactic category system. This approximates the hypothesis of DST that 'macrolevel' symbolic representation is composed at a 'microlevel' of continuously variable levels at activity in the self-attention mechanism. There are still extant theoretical issues posed by MMM, like how to replicate formal contiguity restrictions or the naturalistic learning scenario assumed in MMM using Generative Spoken Language Modelling (Lakhotia, Kharitonov, Hsu, Adi, Polyak, Bolte, Nguyen, Copet, Baevski, Mohamed & Dupoux 2021) to mirror Saussurean sound-meaning mappings. Top-down language modelling clearly opens a number of productive avenues for improving the syntactic-capabilities of Transformers.

5.3 Improving Typological Performance

5.3.1 Against Syntactic Transfer

State-of-the-art techniques used for Transfer Learning, typically evaluated on NLU tasks, do not operate robustly in cross-lingual syntactic transfer– motivating a partition between syntactic and semantic typology in LMs. While our empirical analysis of structural generalisation only focussed on English due to limited evaluation resources, we can apply the typological proposals of the MMM models to multilingual language modelling. I offer a linguistically-motivated analysis of the 'curse of multilinguality' problem (Pfeiffer 2023) in cross-lingual syntactic transfer, where per-language performance drops as models cover more languages.

While mBERT has been very successful for transfer learning when assessed on downstream Natural Language Understanding tasks (Conneau, Wu, Li, Zettlemoyer & Stoyanov 2020), there are two reasons that suggest cross-lingual syntactic transfer is not as effective. Guarasci, Silvestri, De Pietro, Fujita & Esposito (2022) apply the structural probe to analyse syntactic transfer between consistent-null-subject language (CNSL) Italian and non-null-subject (NSL) English and French, suggesting that the probe could reconstruct the dependency parse trees of the Italian sentences without the subject when trained in English/French. However, the observed poor performance of transfer between non-genetically related Italian-English/ English-Italian dyads is also indicative of the failure of robust transfer between non-NSL and CNSL languages. The structural probe faces empirical limitations for measuring syntactic transfer, as mBERT's representations are known not to encode subjecthood purely syntactically– it is modulated by continuous and dependent on semantic and discourse factors (Papadimitriou, Chi, Futrell & Mahowald 2021) which Guarasci et al. do not control for.

Secondly, typological variation falls out of the early sensitivity to initial conditions which creates a language-specific curriculum of successive NO $>$ ALL $>$ SOME learning dynamics in the MMM-model. This cannot be transferred. The poor transfer observed in Guarasci is consistent with a neo-emergent characterisation of null-subject typology, where learners follow a NO (non-NSL) $>$ ALL $>$ SOME (CNSL) parametric pathway (Biberauer 2018). Cross-lingual syntactic transfer of typological features from a shared 'universal' pool is only possible under the Parameter Hierarchies of Roberts (2019), which assumes a one-time selection of FFs as assumed by Chomsky (2000). However, this is not consistent with the Sensitivity Conjecture and the inductive biases of the Transformer supporting emergent syntax. Robert's account is also computationally-implementable due to the absence of an intensional characterisation of what this universal set of FFs includes. Neo-emergence offers an upper bound on the 'shallowness' of transfer-based approaches, which we formulate in (6) below:

(6) ***Shallow Syntactic Transfer Hypothesis (SSTH)****: Full Syntactic Transfer requires a universal 'one-time selection' of FFs, which is (a) not computable and (b) is not possible under the attested MMM-based pre-training dynamic. Full Syntactic Transfer is, therefore, not possible in Transformer-based LMs.*

5.3.2 Meta-Learning, Modularity and Typological Datasets

SSTH motivates a re-appraisal of existing techniques– which are not following the 'right hills' (Bender & Koller 2020) for syntactic transfer. State-of-the-Art techniques like meta-learning and modular pre-training can potentially be repurposed to enable the fast adaptation of Transformers to a low-resource setting to replicate robust very-early Parameter Setting of word order. However, once the 'emergent engine' of the Transformer has kicked-off, contextual restrictions and subsequent Parametric variation will be unable to be transferred.

State-of-the-art techniques like (1) '*modular' adapters*, which adapt multilingual LMs towards the properties of target languages to avoid 'catastrophic forgetting' of emergent knowledge (Ansell, Ponti, Korhonen & Vulić 2022); (2) *clustering* languages during Transfer (Maurya & Desarkar 2022, Choenni & Shutova 2022); and (3) dynamic *language-specific subnetwork*s that split off from the multilingual LM to transfer features between typologically-related languages (Xu, Gui, Ma, Zhang, Ye, Zhang & Huang 2022) are unable to robustly participate in syntactic transfer. This is because they all assume, *contra* the SSTH, that cross-lingual transfer of syntax can happen during pretraining. The SSTH does not impinge on the empirical success of these techniques semantic transfer– instead, it highlights a strict partition between how computational linguists between NLU tasks and structural generalisation, which is expected from a neo-emergent DST perspective where symbolic computation is dissociated from the acquisitional dynamical system and intermediate Conceptual Spaces.

This assessment is supported by Blevins, Gonen & Zettlemoyer (2022) who find that the points of pre-training when models learn to transfer vary cross-linguistically across language pairs and that the layer of multilingual-Transformers exhibit long-term performance degradation as linguistic knowledge propagates to lower layers. Transfer also leads to typologically-intuitive results: Winata, Wu, Kulkarni, Solorio & Preotiuc-Pietro (2022) find that mixture of random source languages is more effective than transferring to unseen typologically-similar languages

A meta-learning algorithm called *Model-Agnostic Meta-learning (MAML)* (Finn, Abbeel & Levine 2017) extracts knowledge from observed tasks to enable a Transformer to adapt to a new data-limited meta-testing task by maximising the 'responsiveness' of loss functions to the new task. While this has been widely-applied in Transfer Learning to learn low-resource languages in a sample-efficient manner, Ponti, Aralikatte, Shrivastava, Reddy & Søgaard (2021) establishes that MAML is ill-suited for cross-lingual NLP as it makes an independent and identically distributed assumption that assumes that evaluation languages share an identical distribution to the source languages. MAML Transfer relies on the evaluation languages being identically distributed. This assumption is incongruous with cross-lingual transfer in realistic scenarios.

Multilingual probing studies typically use coarse-grained datasets, like WALS for probing emergent syntax (Stanczak, Ponti, Torroba Hennigen, Cotterell & Augenstein 2022) and coarse-grained part-of-speech inventories, like UPOS. However, WALS is not clearly related to the patterns of emergence underpinning typological patterns. MMM highlights the benefits of evaluating emergent syntax in Transformers along different levels of granularity. This motivates the creation of neo-emergent typological datasets. Generative linguists have developed a method called the *Parametric Comparison Method (PCM)* which has developed typological datasets in terms of Generative syntactic parameters for diachronic purposes (Marcolli 2016, Ceolin, Guardiano, Irimia & Longobardi 2020). The increased linguistic granularity of PCM-style datasets could be potentially relevant for developing resources for evaluating the typological-uniformity of structural generalisation in Transformers.

The MMM-perspective suggests that meta-learning techniques can be effectively repurposed for better syntactic generalisation. Pfeiffer, Goyal, Lin, Li, Cross, Riedel & Artetxe (2022) pre-train the transformer with modular units 'from the get-go', preparing the model to be extended with additional modular units later on. Langedijk, Dankers, Lippe, Bos, Cardenas Guevara, Yannakoudakis & Shutova (2022) find that integrating MAML with Transformer-based dependency parser, which projects mBERT embeddings through a graph-based bi-affine attention classifier to produce a probability distribution of arc heads for each word, can significantly improve the performance of language transfer and standard supervised learning baselines for a variety of unseen, typologically diverse, and low-resource languages, in a few-shot learning setup. Ponti et al. (2021) also modify the training objectives to account for this limitation. MAML has also been repurposed to improve compositionally in Transformers (Conklin et al. 2021: i.a.). These are all productive avenues to improve syntactic generalisation by augmenting Huebner-style pre-training in monolingual LMs, rather than using multilingual LMs for cross-lingual syntactic transfer.

Leveraging the typological predictions of MMM formulates a theoretically-precise typological upper-bound on current techniques and motivates a partition between cross-lingual syntax and semantics. This top-down approach makes concrete practical recommendations about typological datasets and repurposing existing techniques more effectively.

Future Directions: Multimodal Language Modelling

Salhan, Liu & Collier (2022/in preparation) find that image-text LM CLIP (Radford et al. 2021) performs significantly worse in 'grounding' (associating a text-only embedding with a non-textual conceptual representation) predicate elements than nominal elements (see Figure 21) and that performance on semantic evaluation datasets is better using non-English monolingual LMs, like CLIP-Italian (Bianchi, Attanasio, Pisoni, Terragni, Sarti & Lakshmi 2021) than multilingual CLIP. Model performance was not improved by changing the grounding strategy, such as using a video-text model Video-CLIP (Xu, Ghosh, Huang, Okhonko, Aghajanyan, Metze, Zettlemoyer & Feichtenhofer 2021).

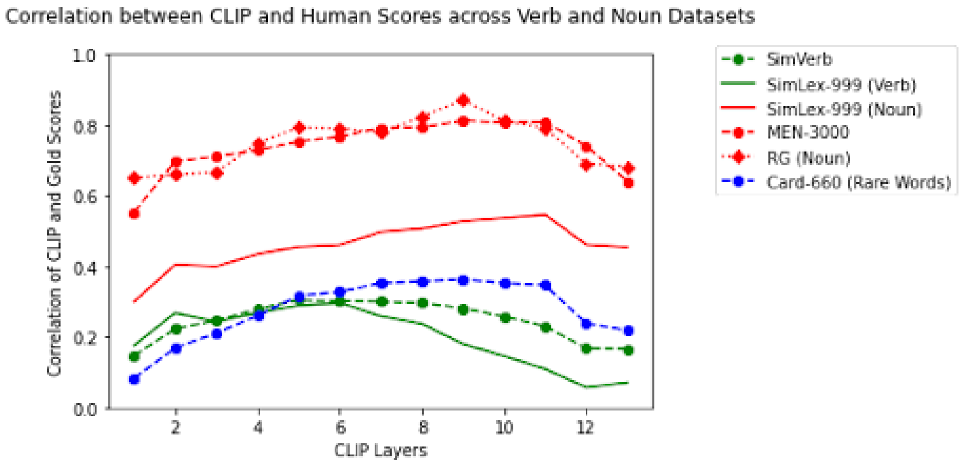


Figure 21 Dissociation of Grounding Nouns and Verbs can be alleviated using by implementing a Sprouting Strategy. Figure from Salhan et al. (2022/in preparation). See Appendix for datasets and code from this experiment.

The key issue for multimodal LMs seems to be category distillation. Cross-Modal Attention learns the visual grounding of NPs into objects and higher semantic information about spatial relations (Ilinykh & Dobnik 2022). Still, these results indicate that multimodal models struggle with attending to and grounding verbal elements. Improving the robust structural generalisation using a Sprouting-inspired pre-training strategy as outlined above can potentially facilitate the development of grounding strategies for event semantics.

6 Conclusions and Further Directions

Pace Vaswani et al. (2017), 'Attention is *not* All You Need' to approximate human emergent syntactic capabilities resource-efficiently across languages. Computational Linguistics has a rich intellectual history that has weaved together rationalism and empiricism and oscillated between symbolic and statistical paradigms throughout the last sixty years (Church 2011, Church & Liberman 2021). While contemporary symbolic approaches in Chomskyan Generativism and connectionist approaches in computational linguistics seem mutually incompatible, neo-emergentist MMM and DST provide computational and theoretical linguists with an 'in-way' to reconcile differing background assumptions.

Linguistic Theory provides an abstract characterisation of the invariant and language-specific substance of grammars that arise through the dynamics of emergence. The Sensitivity Conjecture (section 2.3) highlights an implicit convergence in the inductive biases that Transformers and human learners recruit to support emergent syntax. However, the probing study in section 4.2 highlights a fundamental limitation in the structural generalisation of Transformers. This theoretical perspective advocates a return to the 'early days of Turing, Minsky, Simon and McCarthy' (Chomsky 2022: 364), where computational and theoretical linguistics symbiotically influenced each other.

Understanding how emergence, meta-learning, inductive biases and symbolic supervision are conceived in linguistic theory and are implemented in computational linguistics raises theoretical questions about how symbolic computation can emerge through connectionist neural-architectures. The insight that Transformers simulate the effects of cognitive dynamical systems is a valuable source of external evidence for Dynamical Systems Theory.

Conversely, MMM guides language modelling down a syntactically-motivated cline. It isolates a narrow-set of predictions and practical recommendations: (1) complexity-driven curriculum learning with the syntactic distillation of embeddings; (2) the SSTH places an upper-bound syntactically-naive cross-lingual transfer; (3) repurposing meta-learning algorithms for fast adaption in low-resource settings. I hope to practically-implement these recommendations to improve the structural generalisation of Transformers.

The potential of 'Maximising Minimal Means' in Transformer Language Models will lead to better-performing language models– in an era where 'fair NLP' is of increasing social importance– and can contribute to a reappraisal of how the relationship between connectionism and symbolic representation is viewed in computational and theoretical linguistic inquiry.

References

Acuña-Fariña, J. C. 2012. Agreement, Attraction and Architectural Opportunism. *Journal of Linguistics* 48(2). 257–295. doi:10.1017/S0022226712000084.

Ansell, A., E. Ponti, A. Korhonen & I. Vulić. 2022. Composable Sparse Fine-Tuning for Cross-Lingual Transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1778–1796. Dublin, Ireland: Association for Computational Linguistics. https://aclanthology.org/2022.acl-long.125.

Arora, A., C. Meister & R. Cotterell. 2022. Estimating the Entropy of Linguistic Distributions. In *Proceedings of the 60th Annual Meeting of the Association for*

Computational Linguistics (Volume 2: Short Papers), 175–195. Dublin, Ireland: Association for Computational Linguistics. doi:10.18653/v1/2022.acl-short.20.

Avcu, E. & R. Rhodes. 2022. Experimental Linguistics: Bridging Subregular Linguistics and Cognitive Neuroscience. *Theoretical Linguistics* 48(3-4). 185–198. doi:https://doi.org/10.1515/tl-2022-2038.

Bahdanau, D., K. Cho & Y. Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv preprint arXiv:1409.0473* doi:https://doi.org/10.48550/arXiv.1409.0473.

Bai, J., Y. Wang, Y. Chen, Y. Yang, J. Bai, J. Yu & Y. Tong. 2021. Syntax-BERT: Improving Pre-trained Transformers with Syntax Trees. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 3011–3020. Online: Association for Computational Linguistics. https://aclanthology.org/2021.eacl-main.262.

Bailly, R. & K. Gábor. 2020. Emergence of Syntax Needs Minimal Supervision. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 477–487. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.46.

Baker, M. C. 2008. The Macroparameter in a Microparametric World. *The Limits of Syntactic Variation* 132.

Belinkov, Y. 2022. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics* 48(1). 207–219. https://aclanthology.org/2022.cl-1.7.

Bender, E. M. & A. Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. doi:10.18653/v1/2020.acl-main.463.

Bhatia, S. & B. Dillon. 2022. Processing Agreement in Hindi: When Agreement Feeds Attraction. *Journal of Memory and Language* 125. 104322. doi:https://doi.org/10.1016/j.jml.2022.104322.

Bhattamishra, S., A. Patel, V. Kanade & P. Blunsom. 2022. Simplicity Bias in Transformers and their Ability to Learn Sparse Boolean Functions. *arXiv preprint arXiv:2211.12316* doi:https://doi.org/10.48550/arXiv.2211.12316.

Bianchi, F., G. Attanasio, R. Pisoni, S. Terragni, G. Sarti & S. Lakshmi. 2021. Contrastive Language-Image Pre-Training for the Italian Language. *arXiv preprint arXiv:2108.08688* doi:https://doi.org/10.48550/arXiv.2108.08688.

Biberauer, T. 2011. In Defence of Lexico-Centric Parametric Variation: Two 3rd Factor-Constrained Case Studies. *Paper presented at the Workshop on Formal Grammar and Syntactic Variation: Rethinking Parameters (Madrid).* .

Biberauer, T. 2015. The Limits of Syntactic Variation: An Emergentist Generative Perspective. In *Invited talk given at the Workshop on Language Variation and Change and Cultural Evolution (Centre for Linguistics, History and Diversity, York University, 13 february 2015)*, https://www.york.ac.uk/media/languageandlinguistics/documents/{C}onferences/wlvcce2015/Biberauer_slides.pdf.

Biberauer, T. 2017. Peripheral Significance: A Phasal Perspective on the Grammaticalisation of Speaker Perspective. *Jung* 93.

Biberauer, T. 2018. Pro-drop and Emergent Parameter Hierarchies. In *Null Subjects in Generative Grammar: A Synchronic and Diachronic Perspective*, Oxford University Press. doi:10.1093/oso/9780198815853.003.0005.

Biberauer, T. 2019a. Children Always Go Beyond the Input: The Maximise Minimal Means perspective. *Theoretical Linguistics* 45(3-4). 211–224. doi:https://doi.org/10.1515/tl-2019-0013.

Biberauer, T. 2019b. Factors 2 and 3: Towards a Principled Approach. *Catalan Journal of Linguistics* 45–88. doi:10.5565/rev/catjl.219.

Biberauer, T. & N. Bosch. 2021. The case for Maximise Minimal Means: Linguistic and General-Cognitive Perspectives. *Department of Theoretical and Applied Linguistics, University of Cambridge, [talk presented at SyntaxLab 16/11/21]* .

Biberauer, T., A. Holmberg & I. Roberts. 2014a. A Syntactic Universal and its Consequences. *Linguistic Inquiry* 45(2). 169–225.

Biberauer, T., A. Holmberg, I. Roberts & M. Sheehan. 2014b. Complexity in Comparative Syntax: The View from Modern Parametric Theory. *Measuring Grammatical Complexity* 103–127. doi:https://doi.org/10.1093/acprof:oso/9780199685301.003.0006.

Blevins, T., H. Gonen & L. Zettlemoyer. 2022. Analyzing the Mono- and Cross-Lingual Pretraining Dynamics of Multilingual Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 3575–3590. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. https://aclanthology.org/2022.emnlp-main.234.

Bosch, N. 2022. Emergence, Complexity and Developing grammars: A reinterpretation from a Dynamical Systems perspective. *Cambridge Occasional Papers in Linguistics* https://www.mmll.cam.ac.uk/files/copil_14a_1_bosch.pdf.

Bosch, N. 2023. *Emergent Syntax and Maturation: Rethinking Maturation in Acquisition from a Neo-emergentist lens [MPhil Thesis.* Department of Theoretical and Applied Linguistics, University of Cambridge, presented at SyntaxLab 31/01/23] MA thesis.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al. 2020. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems* 33. 1877–1901. https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Bryant, C., M. Felice, Ø. E. Andersen & T. Briscoe. 2019. The BEA-2019 Shared Task on Grammatical Error Correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 52–75. Florence, Italy: Association for Computational Linguistics. https://aclanthology.org/W19-4406.

Bryant, C., M. Felice & T. Briscoe. 2017. Automatic Annotation and Evaluation of Error Types for Grammatical Error Correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 793–805. Vancouver, Canada: Association for Computational

Linguistics. doi:10.18653/v1/P17-1074.

Ceolin, A., C. Guardiano, M. A. Irimia & G. Longobardi. 2020. Formal Syntax and Deep History. *Frontiers in Psychology* 11. 488871. doi:https://doi.org/10.3389/fpsyg.2020.488871.

Chandlee, J., R. Eyraud & J. Heinz. 2014. Learning Strictly Local Subsequential Functions. *Transactions of the Association for Computational Linguistics* 2. 491–504. doi:10.1162/tacl_a_00198.

Chaves, R. P. & M. T. Putnam. 2022. Islands, Expressiveness, and the Theory/Formalism Confusion. *Theoretical Linguistics* 48(3-4). 219–231. doi:https://doi.org/10.1515/tl-2022-2041.

Chi, E. A., J. Hewitt & C. D. Manning. 2020. Finding Universal Grammatical Relations in Multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5564–5577. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.493.

Chiang, D., P. Cholak & A. Pillay. 2023. Tighter Bounds on the Expressivity of Transformer Encoders. *arXiv preprint arXiv:2301.10743* .

Choenni, R. & E. Shutova. 2022. Investigating Language Relationships in Multilingual Sentence Encoders through the lens of Linguistic Typology. *Computational Linguistics* 48(3). 635–672. doi:10.1162/coli_a_00444.

Chomsky, N. 1956. Three Models for the Description of Language. *IRE Transactions on Information Theory* 2(3). 113–124.

Chomsky, N. 2000. Minimalist Inquiries: The Framework (mitopl 15). *Step by Step: Essays on Minimalist syntax in Honor of Howard Lasnik* 89–155.

Chomsky, N. 2001. Derivation by Phase. *Ken Hale: A Life in Language* .

Chomsky, N. 2005. Three Factors in Language design. *Linguistic Inquiry* 36(1). 1–22.

Chomsky, N. 2022. Genuine Explanation and the Strong Minimalist Thesis. *Cognitive Semantics* 8(3). 347–365.

Chomsky, N. & M. Halle. 1968. *The Sound Pattern of English.* New York: Harper and Row.

Church, K. 2011. A Pendulum Swung Too Far. *Linguistic Issues in Language Technology* 6.

Church, K. & M. Liberman. 2021. The Future of Computational Linguistics: On Beyond Alchemy. *Frontiers in Artificial Intelligence* 4. 625341.

Clark, K., U. Khandelwal, O. Levy & C. D. Manning. 2019. What does BERT look at? an Analysis of BERT's Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 276–286. Florence, Italy: Association for Computational Linguistics. doi:10.18653/v1/W19-4828.

Clark, K., M.-T. Luong, Q. V. Le & C. D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators rather than Generators. *arXiv preprint arXiv:2003.10555* doi:https://doi.org/10.48550/arXiv.2003.10555.

Coecke, B., M. Sadrzadeh & S. Clark. 2010. Mathematical foundations for a compositional distributional model of meaning. *arXiv preprint arXiv:1003.4394* doi:https://doi.org/10.48550/arXiv.1003.439.

Collier, N. H., F. Liu & E. Shareghi. 2022. On Reality and the Limits of Language Data. *arXiv preprint arXiv:2208.11981* doi:https://doi.org/10.48550/arXiv.2208.11981.

Conklin, H., B. Wang, K. Smith & I. Titov. 2021. Meta-learning to Compositionally Generalize. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3322–3335. Online: Association for Computational Linguistics. doi:10.18653/v1/2021.acl-long.258.

Conneau, A., S. Wu, H. Li, L. Zettlemoyer & V. Stoyanov. 2020. Emerging Cross-lingual Structure in Pretrained Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6022–6034. Online: Association for Computational Linguistics. https://aclanthology.org/2020.acl-main.536.

Dai, Z., Z. Yang, Y. Yang, J. Carbonell, Q. Le & R. Salakhutdinov. 2019. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2978–2988. Florence, Italy: Association for Computational Linguistics. doi:10.18653/v1/P19-1285.

Danis, N. & A. Jardine. 2019. Q-theory Representations are Logically Equivalent to Autosegmental Representations. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, 29–38. doi:10.7275/tvj1-k306.

Davis, C., C. Bryant, A. Caines, M. Rei & P. Buttery. 2022. Probing for Targeted Syntactic Knowledge through Grammatical Error Detection. In *Proceedings of the 26th Conference on computational natural language learning (conll)*, 360–373. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics. https://aclanthology.org/2022.conll-1.25.

Devlin, J., M.-W. Chang, K. Lee & K. Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics. doi:10.18653/v1/N19-1423.

Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly et al. 2020. An Image is Worth 16x16 words: Transformers for Image Recognition at Scale. *arXiv preprint arXiv:2010.11929* doi:https://doi.org/10.48550/arXiv.2010.11929.

Dutta, S., T. Gautam, S. Chakrabarti & T. Chakraborty. 2021. Redesigning the Transformer Architecture with Insights from Multi-Particle Dynamical Systems. *Advances in Neural Information Processing Systems* 34. 5531–5544. https://openreview.net/forum?id=e2gqGkFjDHg.

Edelman, B. L., S. Goel, S. Kakade & C. Zhang. 2022. Inductive Biases and Variable Creation in Self-Attention Mechanisms. In *International Conference on Machine Learning*, 5793–5831. PMLR. https://proceedings.mlr.press/v162/edelman22a/edelman22a.pdf.

Eisenstein, J. 2022. Informativeness and Invariance: Two Perspectives on Spurious Correlations in Natural Language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4326–4331. Seattle, United States: Association for

Computational Linguistics. https://aclanthology.org/2022.naacl-main.321.

Emerson, G. 2018. *Functional Distributional Semantics: Learning Linguistically Informed Representations from a Precisely Annotated Corpus*: University of Cambridge dissertation.

Emerson, G. 2020a. Linguists Who Use Probabilistic Models Love Them: Quantification in Functional Distributional Semantics. In *Proceedings of the Probability and Meaning Conference (PaM 2020)*, 41–52. Gothenburg: Association for Computational Linguistics. https://aclanthology.org/2020.pam-1.6.

Emerson, G. 2020b. What are the Goals of Distributional Semantics? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7436–7453. Online: Association for Computational Linguistics. https://aclanthology.org/2020.acl-main.663.

Finlayson, M., A. Mueller, S. Gehrmann, S. Shieber, T. Linzen & Y. Belinkov. 2021. Causal Analysis of Syntactic Agreement Mechanisms in Neural Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1828–1843. Online: Association for Computational Linguistics. doi:10.18653/v1/2021.acl-long.144. https://aclanthology.org/2021.acl-long.144.

Finn, C., P. Abbeel & S. Levine. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *International Conference on Machine Learning*, 1126–1135. PMLR. https://proceedings.mlr.press/v70/finn17a.html.

Firth, J. R. 1957. A Synopsis of Linguistic Theory, 1930-1955. *Studies in Linguistic Analysis* .

Franck, J., F. Sadri Mirdamadi & A. Kahnemuyipour. 2020. Object Attraction and The Role of Structural Hierarchy: Evidence from Persian. *Glossa: a Journal of General Linguistics* 5(1). 27. doi:https://doi.org/10.5334/gjgl.804.

Friedmann, N., A. Belletti & L. Rizzi. 2021. Growing Trees: The Acquisition of the Left Periphery. *Glossa: a Journal of General Linguistics* 6(1).

Futrell, R. & R. Levy. 2017. Noisy-Context Surprisal as a Human Sentence Processing Cost Model. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume, Long Papers*, 688–698.

Futrell, R., E. Wilcox, T. Morita, P. Qian, M. Ballesteros & R. Levy. 2019. Neural Language Models as Psycholinguistic Subjects: Representations of Syntactic State. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume (Long and Short Papers)*, 32–42.

Gong, H., S. Bhat & P. Viswanath. 2018. Embedding Syntax and Semantics of Prepositions via Tensor Decomposition. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume (Long Papers)*, 896–906. New Orleans, Louisiana: Association for Computational Linguistics. doi:10.18653/v1/N18-1082.

Guarasci, R., S. Silvestri, G. De Pietro, H. Fujita & M. Esposito. 2022. Bert Syntactic Transfer: A Computational Experiment on Italian, French and English languages. *Computer Speech & Language* 71. 101261.

Hahn, M. 2020. Theoretical Limitations of Self-Attention in Neural Sequence Models. *Transactions of the Association for Computational Linguistics* 8. 156–171. doi:10.1162/tacl_a_00306.

Hahn, M., J. Degen & R. Futrell. 2021a. Modeling Word and Morpheme Order in Natural Language as an Efficient Trade-Off of Memory and Surprisal. *Psychological Review* 128(4). 726. doi:https://doi.org/10.1037/rev0000269.

Hahn, M., D. Jurafsky & R. Futrell. 2021b. Sensitivity as a Complexity Measure for Sequence Classification Tasks. *Transactions of the Association for Computational Linguistics* 9. 891–908. doi:https://doi.org/10.1162/tacl_a_00403.

Hao, Y., D. Angluin & R. Frank. 2022. Formal Language Recognition by Hard Attention Transformers: Perspectives from Circuit Complexity. *Transactions of the Association for Computational Linguistics* 10. 800–810. doi:https://doi.org/10.1162/tacl_a_00490.

Harnad, S. 1990. The Symbol Grounding Problem. *Physica D: Nonlinear Phenomena* 42(1-3). 335–346.

Heinz, J. 2018. The Computational Nature of Phonological Generalizations. *Phonological Typology, Phonetics and Phonology* 126–195. doi:https://doi.org/10.1515/9783110451931-005.

Hewitt, J. & C. D. Manning. 2019. A Structural Probe for Finding Syntax in Word Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume (Long and Short Papers)*, 4129–4138. Minneapolis, Minnesota: Association for Computational Linguistics. doi:10.18653/v1/N19-1419.

Hu, J., J. Gauthier, P. Qian, E. Wilcox & R. Levy. 2020. A Systematic Assessment of Syntactic Generalization in Neural Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1725–1744. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.158.

Huebner, P. A., E. Sulem, F. Cynthia & D. Roth. 2021. BabyBERTa: Learning More Grammar With Small-Scale Child-Directed Language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, 624–646. Online: Association for Computational Linguistics. doi:10.18653/v1/2021.conll-1.49.

Ilinykh, N. & S. Dobnik. 2022. Attention as Grounding: Exploring Textual and Cross-Modal Attention on Entities and Relations in Language-and-Vision Transformer. In *Findings of the Association for Computational Linguistics: ACL 2022*, 4062–4073. Dublin, Ireland: Association for Computational Linguistics. doi:10.18653/v1/2022.findings-acl.320.

Immer, A., L. Torroba Hennigen, V. Fortuin & R. Cotterell. 2022. Probing as Quantifying Inductive Bias. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1839–1851. Dublin, Ireland: Association for Computational Linguistics. doi:10.18653/v1/2022.acl-long.129.

Inoue, K., S. Ohara, Y. Kuniyoshi & K. Nakajima. 2022. Transient Chaos in Bidirectional Encoder Representations from Transformers. *Physical Review Research* 4(1). 013204. doi:10.1103/PhysRevResearch.4.013204.

Joshi, A. K. 1985. Tree Adjoining Grammars: How much Context-Sensitivity is required to provide reasonable Structural Descriptions? In D. R. Dowty, L. Karttunen & A. M. Zwicky (eds.), *Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives* Studies in Natural Language Processing, 206–250. Cambridge University Press. doi:10.1017/CBO9780511597855.007.

Joshi, M., D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer & O. Levy. 2020. SpanBERT: Improving Pre-Training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics* 8. 64–77. doi:https://doi.org/10.1162/tacl_a_00300.

Kalemaj, I. 2020. Measuring the Complexity of Boolean Functions and the Sensitivity Conjecture. https://cs-people.bu.edu/ikalemaj/media/sensitivity.pdf.

Kaplan, R. M. & M. Kay. 1994. Regular Models of Phonological Rule Systems. *Computational linguistics* 20(3). 331–378. doi:https://dl.acm.org/doi/10.5555/204915.204917.

Lakhotia, K., E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed & E. Dupoux. 2021. On Generative Spoken Language Modeling from Raw Audio. *Transactions of the Association for Computational Linguistics* 9. 1336–1354. doi:10.1162/tacl_a_00430.

Lakretz, Y., T. Desbordes, D. Hupkes & S. Dehaene. 2022. Can Transformers Process Recursive Nested Constructions, Like Humans? In *Proceedings of the 29th International Conference on Computational Linguistics*, 3226–3232. Gyeongju, Republic of Korea: International Committee on Computational Linguistics. https://aclanthology.org/2022.coling-1.285.

Lan, Z., M. Chen, S. Goodman, K. Gimpel, P. Sharma & R. Soricut. 2019. ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations. *arXiv preprint arXiv:1909.11942* https://openreview.net/forum?id=H1eA7AEtvS.

Langedijk, A., V. Dankers, P. Lippe, S. Bos, B. Cardenas Guevara, H. Yannakoudakis & E. Shutova. 2022. Meta-Learning for Fast Cross-Lingual Adaptation in Dependency Parsing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8503–8520. Dublin, Ireland: Association for Computational Linguistics. doi:10.18653/v1/2022.acl-long.582.

Levy, R. 2018. Communicative Efficiency, Uniform Information Density, and the Rational Speech Act Theory. In *Cogsci*, .

Li, B., T. Kim, R. K. Amplayo & F. Keller. 2020. Heads-up! Unsupervised Constituency Parsing via Self-Attention Heads. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 409–424. Suzhou, China: Association for Computational Linguistics. https://aclanthology.org/2020.aACL-main.43.

Limisiewicz, T., D. Mareček & R. Rosa. 2020. Universal Dependencies According to BERT: Both More Specific and More General. In *Findings of the Association* *for Computational Linguistics: Emnlp 2020*, 2710–2722. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.findings-emnlp.245.

Liu, B., J. T. Ash, S. Goel, A. Krishnamurthy & C. Zhang. 2022. Transformers Learn Shortcuts to Automata. *arXiv preprint arXiv:2210.10749* https://openreview.net/forum?id=De4FYqjFueZ.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer & V. Stoyanov. 2019. Roberta: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692* https://openreview.net/forum?id=SyxS0T4tvS.

Manning, C. D. 2015. Computational Linguistics and Deep Learning. *Computational Linguistics* 41(4). 701–707. doi:https://doi.org/10.1162/COLI_a_00239.

Marcolli, M. 2016. Syntactic Parameters and a Coding Theory perspective on Entropy and Complexity of Language Families. *Entropy* 18(4). 110. doi:https://doi.org/10.3390/e18040110.

Marvin, R. & T. Linzen. 2018. Targeted Syntactic Evaluation of Language Models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1192–1202. Brussels, Belgium: Association for Computational Linguistics. doi:10.18653/v1/D18-1151. https://aclanthology.org/D18-1151.

Maudslay, R. H., J. Valvoda, T. Pimentel, A. Williams & R. Cotterell. 2020. A Tale of a Probe and a Parser. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7389–7395. Online: Association for Computational Linguistics. https://aclanthology.org/2020.acl-main.659.

Maurya, K. & M. Desarkar. 2022. Meta-x$_{NLG}$: A Meta-Learning Approach Based on Language Clustering for Zero-Shot Cross-Lingual Transfer and Generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, 269–284. Dublin, Ireland: Association for Computational Linguistics. https://aclanthology.org/2022.findings-acl.24.

Merrill, W., A. Sabharwal & N. A. Smith. 2022. Saturated Transformers are Constant-Depth Threshold Circuits. *Transactions of the Association for Computational Linguistics* 10. 843–856. doi:https://doi.org/10.1162/tacl_a_00493.

Mikolov, T., K. Chen, G. Corrado & J. Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado & J. Dean. 2013b. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems* 26.

Mueller, A., G. Nicolai, P. Petrou-Zeniou, N. Talmina & T. Linzen. 2020. Cross-Linguistic Syntactic Evaluation of Word Prediction Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5523–5539. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.490.

Müller-Eberstein, M., R. van der Goot & B. Plank. 2022a. Probing for Labeled Dependency Trees. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7711–7726. Dublin, Ireland: Association for Computational Linguistics. https://aclanthology.org/2022.acl-long.532.

Müller-Eberstein, M., R. van der Goot & B. Plank. 2022b. Spectral Probing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 7730–7741. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. https://aclanthology.org/2022.emnlp-main.527.

Novak, R., Y. Bahri, D. A. Abolafia, J. Pennington & J. Sohl-Dickstein. 2018. Sensitivity and Generalization in Neural Networks: An Empirical Study. *arXiv preprint arXiv:1802.08760* https://openreview.net/forum?id=HJC2SzZCW.

Papadimitriou, I., E. A. Chi, R. Futrell & K. Mahowald. 2021. Deep Subjecthood: Higher-Order Grammatical Features in Multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2522–2532. Online: Association for Computational Linguistics. https://aclanthology.org/2021.eacl-main.215.

Patel, A., S. Bhattamishra, P. Blunsom & N. Goyal. 2022. Revisiting the Compositional Generalization Abilities of Neural Sequence Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 424–434. Dublin, Ireland: Association for Computational Linguistics. https://aclanthology.org/2022.acl-short.46.

Pérez-Mayos, L., M. Ballesteros & L. Wanner. 2021a. How Much Pretraining Data Do Language Models Need to Learn Syntax? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1571–1582. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. doi:10.18653/v1/2021.emnlp-main.118.

Pérez-Mayos, L., R. Carlini, M. Ballesteros & L. Wanner. 2021b. On the Evolution of Syntactic Information Encoded by BERT's Contextualized Representations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2243–2258. Online: Association for Computational Linguistics. doi:10.18653/v1/2021.eacl-main.191.

Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee & L. Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume (Long Papers)*, 2227–2237. New Orleans, Louisiana: Association for Computational Linguistics. doi:10.18653/v1/N18-1202.

Pfeiffer, J. 2023. Modular and Compositional Transfer Learning. Natural Language and Information Processing (NLIP) Seminar Series [24th February]. https://www.talks.cam.ac.uk/talk/index/196519.

Pfeiffer, J., N. Goyal, X. Lin, X. Li, J. Cross, S. Riedel & M. Artetxe. 2022. Lifting the Curse of Multilinguality by Pre-Training Modular Transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3479–3495. Seattle, United States: Association for Computational Linguistics. https://aclanthology.org/2022.naacl-main.255.

Ponti, E. 2021. Inductive Bias and Modular Design for Sample-Efficient Neural Language Learning.

Ponti, E. M., R. Aralikatte, D. Shrivastava, S. Reddy & A. Søgaard. 2021. Minimax and Neyman–Pearson Meta-Learning for Outlier Languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 1245–1260. Online: Association for Computational Linguistics. doi:10.18653/v1/2021.findings-acl.106.

Qian, P., T. Naseem, R. Levy & R. Fernandez Astudillo. 2021. Structural Guidance for Transformer Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3735–3745. Online: Association for Computational Linguistics. https://aclanthology.org/2021.acl-long.289.

Radford, A., J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark et al. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR. http://proceedings.mlr.press/v139/radford21a.

Radford, A., K. Narasimhan, T. Salimans, I. Sutskever et al. 2018. Improving Language Understanding by Generative Pre-Training. https://paperswithcode.com/paper/improving-language-understanding-by.

Ramchand, G. & P. Svenonius. 2014. Deriving the Functional Hierarchy. *Language Sciences* 46. 152–174.

Ravfogel, S., Y. Elazar, J. Goldberger & Y. Goldberg. 2020. Unsupervised Distillation of Syntactic Information from Contextualized Word Representations. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 91–106. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.BlackboxNLP-1.9.

Ravishankar, V., A. Kulmizev, M. Abdou, A. Søgaard & J. Nivre. 2021. Attention Can Reflect Syntactic Structure (If You Let It). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 3031–3045. Online: Association for Computational Linguistics. doi:10.18653/v1/2021.eacl-main.264.

Rawski, J. & J. Heinz. 2019. No Free Lunch in Linguistics or Machine Learning: Response to Pater. *Language* 95(1). e125–e135.

Rizzi, L. & L. Haegeman. 1997. The Fine Structure of the Left Periphery.

Roberts, I. 2012. A Programme for Comparative Research. *Parameter Theory and Linguistic Change* 2. 320.

Roberts, I. 2019. *Parameter Hierarchies and Universal Grammar*. Oxford University Press.

Salhan, F. Liu & N. Collier. 2022/in preparation. Multimodal Language Modelling across Languages and Cultures: Grounding Strategies for Nouns and Verbs. *Research Project, Language Technology Lab, Department of Theoretical and Applied Linguistics, University of Cambridge [partially supported by a research award from Gonville & Caius College, Cambridge]* .

Sartran, L., S. Barrett, A. Kuncoro, M. Stanojević, P. Blunsom & C. Dyer. 2022. Transformer Grammars: Augmenting Transformer Language Models with Syntactic Inductive Biases at Scale. *Transactions of the Association for Computational*

Linguistics 10. 1423–1439. doi:https://doi.org/10.1162/tacl_a_00526.

Schick, T. & H. Schütze. 2021. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 255–269. Online: Association for Computational Linguistics. doi:10.18653/v1/2021.eacl-main.20.

Schuster, M. & K. Nakajima. 2012. Japanese and Korean Voice Search. In *2012 IEEE International Conference on Acoustics, Speech and Signal processing (ICASSP)*, 5149–5152. IEEE. doi:10.1109/ICASSP.2012.6289079.

Schwartz, R. & G. Stanovsky. 2022. On the Limitations of Dataset Balancing: The Lost Battle Against Spurious Correlations. In *Findings of the Association for Computational Linguistics: naacl 2022*, 2182–2194. Seattle, United States: Association for Computational Linguistics. https://aclanthology.org/2022.findings-naacl.168.

Sennrich, R., B. Haddow & A. Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1715–1725. Berlin, Germany: Association for Computational Linguistics. doi:10.18653/v1/P16-1162.

Sheehan, M., T. Biberauer, I. Roberts & A. Holmberg. 2017. *The Final-over-Final Condition: A Syntactic Universal*, vol. 76. MIT Press.

Stabler, E. P. 2011. Computational Perspectives on Minimalism. In *The Oxford Handbook of Linguistic Minimalism*, Oxford University Press. doi:https://doi.org/10.1093/oxfordhb/9780199549368.013.0027.

Stanczak, K., E. Ponti, L. Torroba Hennigen, R. Cotterell & I. Augenstein. 2022. Same Neurons, Different Languages: Probing Morphosyntax in Multilingual Pre-trained Models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1589–1598. Seattle, United States: Association for Computational Linguistics. doi:10.18653/v1/2022.naacl-main.114.

Surkov, M., V. Mosin & I. Yamshchikov. 2022. Do Data-based Curricula Work? In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, 119–128. Dublin, Ireland: Association for Computational Linguistics. doi:10.18653/v1/2022.insights-1.16.

Tamkin, A., D. Jurafsky & N. Goodman. 2020. Language through a Prism: A Spectral Approach for Multiscale Language Representations. *Advances in Neural Information Processing Systems* 33. 5492–5504. https://proceedings.neurips.cc/paper/2020/hash/3acb2a202ae4bea8840224e6fce16fd0-Abstract.html.

Thrush, T., E. Wilcox & R. Levy. 2020. Investigating Novel Verb Learning in BERT: Selectional Preference Classes and Alternation-Based Syntactic Generalization. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 265–275. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.BlackboxNLP-1.25. https://aclanthology.org/2020.{B}lackbox{NLP}-1.25.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser & I. Polosukhin. 2017. Attention is All You Need. *Advances in Neural Infor-*

mation Processing Systems 30. https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Vig, J., S. Gehrmann, Y. Belinkov, S. Qian, D. Nevo, Y. Singer & S. Shieber. 2020. Investigating Gender Bias in Language Models using Causal Mediation Analysis. *Advances in Neural Information Processing Systems* 33. 12388–12401.

Voita, E., D. Talbot, F. Moiseev, R. Sennrich & I. Titov. 2019. Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5797–5808. Florence, Italy: Association for Computational Linguistics. doi:10.18653/v1/P19-1580.

van der Wal, J. 2022. *A Featural Typology of Bantu Agreement*. Oxford University Press. doi:https://doi.org/10.1093/oso/9780198844280.001.0001.

Warstadt, A., A. Parrish, H. Liu, A. Mohananey, W. Peng, S.-F. Wang & S. R. Bowman. 2020. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics* 8. 377–392. https://aclanthology.org/2020.tacl-1.25.

Wei, J., D. Garrette, T. Linzen & E. Pavlick. 2021. Frequency Effects on Syntactic Rule Learning in Transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 932–948. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. https://aclanthology.org/2021.emnlp-main.72.

Wilcox, E., R. Levy, T. Morita & R. Futrell. 2018. What do RNN Language Models Learn about Filler–Gap Dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 211–221.

Wilcox, E. G., R. Futrell & R. Levy. 2022. Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry* 1–88. doi:https://doi.org/10.1162/ling_a_00491.

Wiltschko, M. 2014. *The Universal Structure of Categories: Towards a Formal Typology*, vol. 142. Cambridge University Press.

Wiltschko, M. 2021. *The Grammar of Interactional Language*. Cambridge University Press.

Winata, G., S. Wu, M. Kulkarni, T. Solorio & D. Preotiuc-Pietro. 2022. Cross-lingual Few-Shot Learning on Unseen Languages. In *Proceedings of the 2nd Conference of the asia-pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 777–791. Online only: Association for Computational Linguistics. https://aclanthology.org/2022.aACL-main.59.

Xu, H., G. Ghosh, P.-Y. Huang, D. Okhonko, A. Aghajanyan, F. Metze, L. Zettlemoyer & C. Feichtenhofer. 2021. VideoCLIP: Contrastive Pre-Training for Zero-Shot Video-Text Understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6787–6800. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. https://aclanthology.org/2021.emnlp-main.544.

Xu, N., T. Gui, R. Ma, Q. Zhang, J. Ye, M. Zhang & X. Huang. 2022. Cross-Linguistic Syntactic Difference in Multilingual BERT: How Good is It and

How Does It Affect Transfer? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 8073–8092. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. https://aclanthology.org/2022.emnlp-main.552.

Yang, Z., Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov & Q. V. Le. 2019. Xl-Net: Generalized Autoregressive Pretraining for Language Understanding. *Advances in Neural Information Processing Systems* 32. https://papers.nips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html.

Yannakoudakis, H., T. Briscoe & B. Medlock. 2011. A New Dataset and Method for Automatically Grading ESOL Texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 180–189. Portland, Oregon, USA: Association for Computational Linguistics. https://aclanthology.org/P11-1019.

Yao, S., B. Peng, C. Papadimitriou & K. Narasimhan. 2021. Self-Attention Networks can process Bounded Hierarchical Languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3770–3785. Online: Association for Computational Linguistics. doi:10.18653/v1/2021.acl-long.292.

Zhang, S., W. Lijie, X. Xiao & H. Wu. 2022. Syntax-Guided Contrastive Learning for Pre-trained Language Model. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2430–2440. Dublin, Ireland: Association for Computational Linguistics. doi:10.18653/v1/2022.findings-acl.191.

Zhang, T. & T. B. Hashimoto. 2021. On the Inductive Bias of Masked Language Modeling: From Statistical to Syntactic Dependencies. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5131–5146. Online: Association for Computational Linguistics. https://aclanthology.org/2021.naacl-main.404.

Zhou, H., Y. Li, Z. Li & M. Zhang. 2022. Bridging Pre-Trained Language Models and Hand-crafted Features for Unsupervised POS Tagging. In *Findings of the Association for Computational Linguistics: ACL 2022*, 3276–3290. Dublin, Ireland: Association for Computational Linguistics. doi:10.18653/v1/2022.findings-acl.259.

Appendix: Statement on Data Availability

All Transformer-based LMs referenced in this dissertation are available open-source from https://huggingface.co/models. All the code and datasets used in this dissertation are available in this GitHub repository: https://github.com/suchirsalhan/Part-IIB-Dissertation.

The repository contains the following:

i. *GED Probing Dataset Sample:* This contains a sample of the Grammatical Error Detection probing results for Layer 12 of BERT. All probing experiments cited were run in the Computer Lab in research supported by Cambridge University

Press & Assessment. The per-layer probing results were personally shared by Chris Davis and Andrew Caines for further analysis in this dissertation. Figures produced were based on F1-macro scores as reported in the data sample. The dataset shared contains a similar set of evaluation scores for all five models for each layer.

ii. The targeted syntactic evaluation dataset developed by Marvin & Linzen is included in the repository and also includes links to the W&I-FCE corpus and the ERRANT annotation toolkit (Bryant et al. 2017). Another folder contains examples of the GED Predictions when evaluated sentences in the dataset according to different agreement attraction stimuli.

iii. Relevant code from the *Causal Mediation Analysis* performed by Finlayson et al. (2021) It was not replicated as this causal mediation analysis is very computationally expensive– the authors note that it 'can take hours on a GPU. The outputs also require gigabytes of space for the largest models.'

iv. A selection of the code from Salhan et al. (2022/in preparation) for intrinsic semantic evaluation of multimodal CLIP is also included.

Suchir Salhan
University of Cambridge
sas245@cam.ac.uk