Synthese (2024) 203:154
https://doi.org/10.1007/s11229-024-04566-3

ORIGINAL RESEARCH

Natural language syntax complies with the free-energy principle

Elliot Murphy[1,2] · Emma Holmes[3,4] · Karl Friston[4]

Received: 9 July 2023 / Accepted: 15 March 2024

Abstract

Natural language syntax yields an unbounded array of hierarchically structured expressions. We claim that these are used in the service of active inference in accord with the free-energy principle (FEP). While conceptual advances alongside modelling and simulation work have attempted to connect speech segmentation and linguistic communication with the FEP, we extend this program to the underlying computations responsible for generating syntactic objects. We argue that recently proposed principles of economy in language design—such as "minimal search" criteria from theoretical syntax—adhere to the FEP. This affords a greater degree of explanatory power to the FEP—with respect to higher language functions—and offers linguistics a grounding in first principles with respect to computability. While we mostly focus on building new principled conceptual relations between syntax and the FEP, we also show through a sample of preliminary examples how both tree-geometric depth and a Kolmogorov complexity estimate (recruiting a Lempel–Ziv compression algorithm) can be used to accurately predict legal operations on syntactic workspaces, directly in line with formulations of variational free energy minimization. This is used to motivate a general principle of language design that we term Turing–Chomsky Compression (TCC). We use TCC to align concerns of linguists with the normative account of self-organization

✉ Emma Holmes
emma.holmes@ucl.ac.uk

Elliot Murphy
elliot.murphy@uth.tmc.edu

[1] Vivian L. Smith Department of Neurosurgery, McGovern Medical School, University of Texas Health Science Center, Houston, TX 77030, USA

[2] Texas Institute for Restorative Neurotechnologies, University of Texas Health Science Center, Houston, TX 77030, USA

[3] Department of Speech Hearing and Phonetic Sciences, University College London, London WC1N 1PF, UK

[4] The Wellcome Centre for Human Neuroimaging, UCL Queen Square Institute of Neurology, London WC1N 3AR, UK

Published online: 03 May 2024

furnished by the FEP, by marshalling evidence from theoretical linguistics and psycholinguistics to ground core principles of efficient syntactic computation within active inference.

Keywords Active inference · Free-energy principle · Language · Lempel–Ziv · Kolmogorov complexity · Compression

Implementational models of language must be plausible from the perspective of neuroanatomy (Embick & Poeppel, 2015), but they must also be plausible from the perspective of how biophysical systems behave. We will argue that the structuring influence of the free-energy principle (FEP) can be detected in language, not only via narrative (Bouizegarene et al., 2020), interpersonal dialogue (Friston et al., 2020), cooperative/intentional communication (Vasil et al., 2020) and speech segmentation (Friston et al., 2021), but also at the more fundamental level of what linguists consider to be basic structure-building computations (Adger Forthcoming, Berwick & Stabler, 2019; Chomsky, 1949, 1951, 1956, 1959, 2021a, b, c, 2023).

Natural language syntax yields an unbounded array of hierarchically structured expressions. We argue that many historical insights into syntax are consistent with the FEP—providing a novel perspective under which the principles governing syntax are not limited to language, but rather reflect domain-general processes. This is consistent with a strain within theoretical linguistics that explores how syntactic computation may adhere to "general principles that may well fall within extra-biological natural law, particularly considerations of minimal computation" (Chomsky, 2011, p. 263), such that certain linguistic theories might be engaging with general properties of organic systems (Chomsky, 2004, 2014). Here, we consider the idea that many aspects of natural language syntax may be special cases of a variational principle of least free-energy. To this end, we examine whether a complexity measure relevant to formulations of free-energy—namely, Kolmogorov complexity (Hutter, 2006; MacKay, 1995; Wallace & Dowe, 1999)—relates to legal syntactic processes.

While the FEP has a substantial explanatory scope, across a large range of cognitive systems, it can also be seen as a method or principle of least action for multidisciplinary research (Ramstead et al., 2021), in much the same way that the notion of economy is typically entertained in linguistics as a *programmatic* notion (Chomsky, 1995). The FEP describes the *optimal* behavior of an organism interacting with the environment. The FEP itself has been argued to be more of a conceptual-mathematical model for self-organizing systems (for some, it is a "generic" model; Barandiaran & Chemero, 2009), or a guiding framework (Colombo & Wright, 2021). Thus, when we argue that natural language syntax "complies" with the FEP, this is not to imply that the FEP necessarily bears any specific, direct predictions for linguistic behaviour. Rather, it motivates the construction of conceptual arguments for how some property of organic systems might be seen as realizing the FEP. Hence, we will mostly focus here on presenting a series of principled conceptual relations between the FEP and natural language syntax, with our goal being to promote more systematic empirical research in the near future, given the space required to fully elaborate conceptual sympathies between two mature scientific fields with extensive histories. Nevertheless, after reviewing some of these general sympathies, we will aim to defend a specific analytic approach to the empirical assessment of syntactic models. We will suggest that syntactic derivations minimising algorithmic complexity are licensed over those that result in structures and derivational paths that are less algorithmically compressible.

We begin by summarising the FEP, and describe how syntactic principles are consistent with it. We consider how the FEP is a variational principle of "least action", such as those that describe systems with conserved quantities (Coopersmith, 2017). We then review key observations from linguistics that speak to the structuring influence of computational efficiency, involving "least effort" and "minimal search" restrictions (Bošković & Lasnik, 2007; Gallego & Martin, 2018; Larson, 2015), viewing language as a product of an individual's mind/brain, following the standard 'I-language' (Chomsky, 1986, 2000) perspective in generative linguistics (i.e., 'internal', 'individual', 'intensional'). After modeling the complexity of postulated minimal search procedures—versus their ungrammatical alternatives across a small but representative number of exemplar cases—we propose a unifying principle for how the goals of the FEP might be realised during the derivation of syntactic structures, which we term Turing–Chomsky Compression (TCC). TCC provides a formal description of how the basic mechanisms of syntax (i.e., the merging of lexical items into hierarchically organized sets) directly comply with the FEP. We conclude by highlighting directions for future work.

1 Active inference and the free-energy principle

Before we evaluate any work pertaining to linguistic behaviour, this section introduces key elements of the FEP that motivate its application to language.

1.1 The free-energy principle

The FEP has a long history in theoretical neuroscience (see Friston, 2010 for a review). It states that any adaptive change in the brain will minimise free-energy, either over evolutionary time or immediate, perceptual time (Ramstead et al., 2018). Free-energy is an information-theoretic quantity and is a function of sensory data and brain states: in brief, it is the upper bound on the 'surprise'—or surprisal (Tribus, 1961)—of sensory data, given predictions that are based on an internal model of how those data were generated. The difference between free-energy and surprise is the difference (specified by the Kullback–Leibler divergence) between probabilistic representations encoded by the brain and the true conditional distribution of the causes of sensory input. This is evident in the following equation, which specifies variational free energy (F) as the negative log probability of observations ($\tilde{o}$) under a generative model (i.e., 'surprise') plus the Kullback–Leibler divergence (D_{KL}) between the approximate posterior distribution and the true posterior distribution (where Q indicates posterior beliefs, $\hat{s}$ indicates the states in the generative model, and$P$ indicates the probability under the internal model):

$$F = -\ln P(\tilde{o}) + D_{KL}[Q(\tilde{s})||P(\tilde{s}|\tilde{o})] \tag{1}$$

Unlike surprise itself, variational free energy can be evaluated (for a detailed explanation, see Friston et al., 2017a). Under simplifying assumptions, free-energy can be considered as the amount of prediction error; for a mathematical comparison, see Friston et al. (2017b). Minimising free energy minimises surprise, and is equivalent to maximising the evidence for the internal model of how sensory data were generated. By minimising free-energy, the brain is essentially performing approximate Bayesian inference. By reformulating variational free energy—in a way that is mathematically equivalent; see Penny et al. (2004), Winn and Bishop (2005)—we see that free-energy can be considered as a trade-off between accuracy and complexity, whereby the best internal model is the one that accurately describes the data in the simplest manner (where E_Q indicates the expected value, and the other variables are the same as those defined above):

$$\begin{aligned} F &= E_Q[\ln Q(\tilde{s}) - \ln P(\tilde{o}|\tilde{s}) - \ln P(\tilde{s})] \\ &= E_Q[\ln Q(\tilde{s}) - \ln P(\tilde{s}|\tilde{o}) - \ln P(\tilde{o})] \\ &= \underbrace{D_{KL}[Q(\tilde{s})||P(\tilde{s}|\tilde{o})]}_{relative\ entropy} - \underbrace{\ln P(\tilde{o})}_{\log\ evidence} \\ &= \underbrace{D_{KL}[Q(\tilde{s})||P(\tilde{s})]}_{complexity} - \underbrace{E_Q[\ln P(\tilde{s}|\tilde{o})]}_{accuracy} - \ln P(\tilde{o})] \end{aligned} \tag{2}$$

Because the Kullback–Leibler divergence can never be less than zero, the variational free energy provides an upper bound on negative log evidence: equivalently, the negative free energy provides a lower bound on log evidence; known as an evidence lower bound (ELBO) in machine learning (Winn & Bishop, 2005). The final equality shows a complementary decomposition of variational free energy into accuracy and complexity. In effect, it reflects the degree of belief updating afforded by some new sensory data; in other words, how much some new evidence causes one to "change one's mind". A good generative model—with the right kind of priors—will minimise the need for extensive belief updating and thereby minimise complexity.

The complexity part of variational free energy will become important later, when we will be evaluating the complexity of syntactic processes using a measure derived both in spirit and mathematical heritage from the foundations of the FEP. To present some initial details about this, consider how complexity also appears in treatments of universal computation (Hutter, 2006) and in the guise of minimum message or description lengths (Wallace & Dowe, 1999). Indeed, in machine learning, variational free energy minimisation has been cast as minimising complexity—or maximising efficiency in this setting (MacKay, 1995). One sees that same theme emerge in predictive coding—and related—formulations of free energy minimisation, where the underlying theme is to compress representations (Schmidhuber, 2010), thereby maximising their efficiency (Barlow, 1961; Linsker, 1990; Rao & Ballard, 1999). We will return to these topics below when we begin to formalise features of natural language syntax.

Lastly, the FEP can also be considered from the perspective of a Markov blanket (see Kirchhoff et al., 2018; Palacios et al., 2020; Parr et al., 2020 for detailed explanation), which instantiates a statistical boundary between internal states and external states. In other words, internal (e.g., in the brain) and external (e.g., in the external world) states are conditionally independent: they can only influence one another through blanket states. The blanket states can be partitioned into sensory states and active states. External states affect internal states only through sensory states, while internal states affect external states only through active states (Murphy, 2023a). The implicit circular causality is formally identical to the perception–action cycle (Fuster, 2004). Under previous accounts (Friston et al., 2017a, 2017b), the brain can minimise free-energy either through perception or action. The former refers to optimising (i.e., using approximate Bayesian inference to invert) its probabilistic generative model that specifies how hidden states cause sensory data; in other words, inferring the cause of sensory consequences by minimising variational free energy. The latter refers to initiating actions to sample data that are predicted by its model—which we turn to next.

1.2 Active inference

The enactive component of active inference rests on the assumption that action is biased to realize preferred outcomes. Beliefs about which actions are best to pursue rely on predictions about future outcomes, and the probability of pursuing any particular outcome is given by the *expected free energy* of that action. Expected free energy (G) can be expressed as the combination of extrinsic and epistemic value (Friston et al., 2017b), where π is a series of actions (i.e., the policy) being pursued, τ is the current time point, and the other variables are the same as those defined above:

$$\begin{aligned} G(\pi) &= \sum_{t} G(\pi, \tau) \\ G(\pi, \tau) &= E_Q[\ln Q(\tilde{s}|\pi) - \ln Q(\tilde{s}_\tau|o_\tau, \pi) - \ln P(\tilde{o}_\tau)] \\ &= \underbrace{E_Q[\ln Q(\tilde{s}|\pi) - \ln Q(\tilde{s}_\tau|o_\tau, \pi)]}_{(negative)\ mutual\ \inf ormation} - \underbrace{E_Q[\ln P(\tilde{o}_\tau)]}_{expected \log\ evidence} \\ &= \underbrace{E_Q[\ln Q(o_\tau|\pi) - \ln Q(o_\tau|\tilde{s}_\tau, \pi)]}_{(negative)\ epistemic\ value} - \underbrace{E_Q[\ln P(\tilde{o}_\tau)]}_{extrinsic\ value} \end{aligned} \tag{3}$$

Extrinsic value is the expected evidence for the internal model under a particular course of action, whereas epistemic value is the expected information gain; in other words, the extent a series of actions reduces uncertainty.

Notice that the expected versions of Kullback–Leibler divergence and log evidence in the free energy A2) now become intrinsic and extrinsic value respectively (Eq. 3). As such, selecting an action to minimise expected free energy reduces expected surprise (i.e., uncertainty) in virtue of maximising the information gain while—at the same time—maximising the expected log evidence; namely, actively self-evidencing (Hohwy, 2016, 2020). When applied to a variety of topics in cognitive neuroscience, active inference has been shown to predict human behaviour and neuronal responses; e.g., Brown et al. (2013), Friston et al., (2017a, 2017b), Friston (2019), Smith et al. (2019).

As will soon become clear, we will be using these observations concerning complexity to motivate a specific application of these ideas to natural language syntax, utilizing a measurement of algorithmic complexity that shares a mathematical heritage with the FEP, as outlined here.

1.3 Belief updating

Belief updating refers to a process by which free energy is minimised. By specifying a process theory that explains neuronal responses during perception and action, neuronal dynamics have previously been cast as a gradient flow on free energy (known as variational free energy in physics, introduced in Feynman, 1972; see Hinton & Zemel, 1994); we refer the reader to Friston et al. (2017b) for a treatment of neuronal message passing and belief propagation. That is to say, any neural process can be formulated as a minimisation of the same quantity used in approximate Bayesian inference (Andrews, 2021; Hohwy, 2016). We provide an example of the computational architecture implied by this formulation of belief updating in the brain (Fig. 1). This illustrative example is based upon a discrete state space generative model, where the equations describe the solutions to Bayesian belief updating of expectations pertaining to hidden states, policies, policy precision and parameters.

In short, the brain seeks to minimise free energy, which is mathematically equivalent to maximising model evidence. This view of neuronal responses can be conceived with respect to Hamilton's principle of least action, whereby action is the path integral of free energy.

1.4 Previous applications

Applying active inference to language relies on finding the right sort of generative model, and many different structures and forms of generative models are possible. Most relevant to the current application, deep temporal models accommodate a nesting of states that unfold at different temporal scales. Since language output is inherently temporal, this leads to the question of how to map hierarchical structures onto serial outputs (Epstein et al., 1998), and models that are deep in time allow us to deconstruct associated nested structures.

Recently, a deep temporal model for communication was developed based on a simulated conversation between two synthetic subjects, showing that certain behavioural and neurophysiological correlates of communication arise under variational message passing (Friston et al., 2020). The model incorporates various levels that operate at different temporal scales. At the lowest level, it specifies mappings among syntactic units that, when combined with particular semantic beliefs, predict individual words. At a higher level (i.e., longer temporal scales), the model contains beliefs about the context it is in, which specifies the syntactic structure at the level below. This model is congruent with core assumptions from linguistics concerning the generative nature of language. Specifically, elementary syntactic units provide *belief structures* that are used to reduce uncertainty about the world, through rapid and reflexive categorization of events, objects and their relations. Then, sentential representations can be thought of

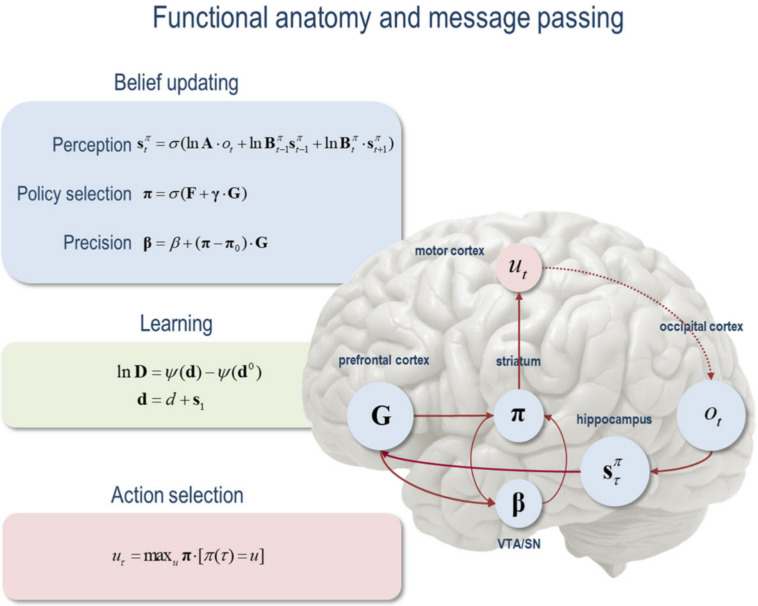
Functional anatomy and message passing
Belief updating
Perception $\mathbf{s}_t^\pi = \sigma(\ln \mathbf{A} \cdot o_t + \ln \mathbf{B}_{t-1}^\pi \mathbf{s}_{t-1}^\pi + \ln \mathbf{B}_t^\pi \cdot \mathbf{s}_{t+1}^\pi)$
Policy selection $\boldsymbol{\pi} = \sigma(\mathbf{F} + \boldsymbol{\gamma} \cdot \mathbf{G})$
Precision $\boldsymbol{\beta} = \beta + (\boldsymbol{\pi} - \boldsymbol{\pi}_0) \cdot \mathbf{G}$
Learning
$\ln \mathbf{D} = \psi(\mathbf{d}) - \psi(\mathbf{d}^0)$
$\mathbf{d} = d + \mathbf{s}_1$
Action selection
$u_\tau = \max_u \boldsymbol{\pi} \cdot [\pi(\tau) = u]$
motor cortex
u_t
occipital cortex
prefrontal cortex
striatum
$\mathbf{G}$
$\boldsymbol{\pi}$
hippocampus
o_t
$\mathbf{s}_\tau^\pi$
$\boldsymbol{\beta}$
VTA/SN

Fig. 1 Schematic overview of belief updates for active inference under discrete Markovian models. The left panel lists the belief updating equations, associating various updates with action, perception, policy selection, precision and learning. The left panel assigns the variables (sufficient statistics or expectations) that are updated to various brain areas. This attribution serves to illustrate a rough functional anatomy—implied by the form of the belief updates. In this simplified scheme, we have assigned observed outcomes to visual representations in the occipital cortex and updates to hidden states to the hippocampal formation. The evaluation of policies, in terms of their (expected) free energy, has been placed in the ventral prefrontal cortex. Expectations about policies per se and the precision of these beliefs have been attributed to striatal and ventral tegmental areas to indicate a putative role for dopamine in encoding precision. Finally, beliefs about policies are used to create Bayesian model averages over future states, which are fulfilled by action. The red arrows denote message passing. In brief, the parameters of the generative model correspond to matrices or arrays encoding the likelihood **A**, prior state transitions **B**, and initial hidden states **D**. **F** corresponds to the free energy of each policy and **G** corresponds to the expected free energy, which is weighted by a precision or softmax parameter $\boldsymbol{\gamma}$ that is usually attributed to dopaminergic neurotransmission. See Friston et al. (2017a) for further explanation of the equations and variables

as structures designed (partially) to consolidate and appropriately frame experiences, and to contextualise and anticipate future experiences. The range of parseable syntactic structures available to comprehenders provides alternate hypotheses that afford parsimonious explanations for sensory data and, as such, preclude overfitting. If the complexities of linguistic stimuli can be efficiently mapped to a small series of regular syntactic formats, this contributes to the brain's goal of restricting itself to a limited number of states. Essentially, the active inference models for linguistic communication previously developed can generally capture interactional dynamics between agents. We take as our point of departure here the question of what needs to happen *within* a single agent's language system. Meanwhile, the psycholinguistic validity and polynomial parseability of minimalist 'bare phrase structure' grammars have recently been demonstrated (Torr et al., 2019), but little else has been said of how to motivate the fundamentals of syntactic theory from extra-linguistic computational constraints.

Before moving forward, we stress here that we will be working within the framework of theoretical linguistics (which deals with *derivational* stages of word-by-word, element-wise operations that underlie sentences), and not a framework such as corpus linguistics (which deals with the *output* of the generative/derivational process). Questions we raise here therefore cannot be addressed by consulting large corpora, but instead require investigation of incremental computational steps that ultimately appear to be responsible for the complex forms of human language behavior studied by sociolinguists, corpus linguists and historical linguists. Relatedly, embracing the traditional distinction between competence and performance, our focus will be on the former (the mental formatting and generation of linguistic structure), and not on the range of complex cognitive processes that enter into the use of language in a specific context.

2 Computational principles and syntactic hierarchies

2.1 A system of discrete infinity

How can the FEP contribute to our understanding of syntactic computation? Most immediately, it provides fundamental constraints on the physical realisation of a computational system. Consider first the three principles in classical recursive function theory which allow functions to compose (Kleene, 1952): substitution; primitive recursion; and minimisation. These are all designed in a way that one might think of as computationally efficient: they reuse the output of earlier computations. Substitution replaces the argument of a function with another function (possibly the original one); primitive recursion defines a new function on the basis of a recursive call to itself, bottoming out in a previously defined function (Lobina, 2017); minimisation (also termed ‘bounded search’) produces the output of a function with the smallest number of steps (see also Piantadosi, 2021, for whom human thought is essentially like Church encoding). More broadly, we note that free energy minimization—by construction—entails Bayesian inference, which in turn is a computational process, and so the FEP entails computationalism (Korbak, 2021) and at least a type of (basic) computational architecture for language we assume here (but see Kirchhoff & Robertson, 2018). Examining some core principles of recursion, natural language clearly exhibits minimisation, while binary branching of structures (Radford, 2016) limits redundant computation, reducing the range of possible computations. Even limitations on short-term memory have been hypothesized to contribute to the efficiency of memory search (MacGregor, 1987).

Syntax involves the construction of binary-branching hierarchically organized sets via the operation MERGE, taking objects from the lexicon or objects already part of the workspace (Marcolli et al., 2023). For example, given the set {X, Y}, we can either select a new lexical object and MERGE it, to form {Z, {X, Y}}, or we can select an existing object and MERGE it to the same workspace, to form {X, {X, Y}}.[1] MERGE serves a similar role to an elementary function, as in the theory of computability (e.g., the zero function, the identity function), in that it is meant to be non-decomposable. Putting many subsidiary technical details aside, these sets are then 'labeled' and given a syntactic identity, or a 'head' (Frampton & Gutmann, 1999; Leivada et al., 2023; Murphy et al., 2022, 2023; Murphy, 2015a, b, c, 2023a; Woolnough et al., 2023), based on which element is most structurally prominent and easiest to search for (i.e., Z in the structure {Z, {X, Y}}).[2] Labeling takes place when conceptual systems access the structures generated by syntax. This occurs at distinct derivational punctuations based on the configuration and subcategorization demands of the lexical items involved (e.g., in many instances subjects seem to be featurally richer than objects, and provide the relevant feature for the label; Longobardi, 2008). For example, in the case of head-complement structures this is done immediately after MERGE (Bošković, 2016). MERGE can also derive some set-theoretic properties of linguistic relations, such as *membership*, *dominate* and *term-of*, as well as the derived relation of *c-command* (= sister of) which is relevant for interpreting hierarchical relations between linguistic elements (Haegeman, 1994). These also appear to be the simplest possible formal relations entertained, potentially indexing a feature of organic computation that adheres closely to criteria of simplicity (Chomsky, 2022).

One might also think of MERGE as physical information coarse-graining (i.e., the removal of superfluous degrees of freedom in order to describe a given physical system at a different scale), with the core process of syntax being information renormalization according to different timescales. For instance, MERGE can be framed as a probability tensor implementing coarse-graining, akin to a probabilistic context-free grammar (Gallego & Orús, 2022). The model proposed in Gallego and Orús (2022, p. 20) assumes that language is "the cheapest non-trivial computational system", exhibiting a high degree of efficiency with respect to its MERGE-based coarse-graining. More recently, MERGE has been described mathematically in terms of Hopf algebras, with a formalism similar to the one arising in the physics of renormalization (Marcolli et al., 2023), and the persistent homology method of topological data analysis and dimensional analysis techniques has been used to study syntactic parameters (Port et al., 2022). Hence, both the computational and mathematical foundations of syntax can be cast in ways that directly accord with the demands of the FEP and active inference.

Natural language syntax exhibits *discrete units* which lead to a *discreteness-continuity duality* (the boundary between syntactic categories can be non-distinct).[3]

[1] There are recent debates concerning whether merging new lexical objects is more computationally demanding (since it requires searching the lexicon) than merging lexical objects that are already in the workspace. This may motivate a Move-over-Merge bias (reversing the Merge-over-Move assumption from the early minimalism of the 1990s), however we leave this issue to one side.

[2] There is increasing evidence that only elements in the workspace that have been labeled can be subject to movement (Bošković 2021).

[3] In active inference, the use of discrete—as opposed to continuous—states in generative models is an enormously potent way of minimising complexity. For example, if it is sufficient to carve the world (i.e., the causes of my sensations) into a small number of hidden states, one can minimise the complexity of belief updating by not redundantly representing all the fine-grained structure within any one state. Similarly, factorisation plays a key role in structuring our hypotheses or expectations that provide the best explanation

Syntax is driven by *closeness of computation* (syntactic objects X and Y form a distinct syntactic object, {X, Y}). Its objects are *bounded* (a fixed list, e.g., N, V, Adj, Adv, P, C, T, *n*, *v*, Asp, Cl, Neg, Q, Det) and their hierarchical ordering is based on a specific functional sequence such as C-T-*v*-V (e.g., C is always higher than V; Starke, 2004) which imposes direct restrictions on combinatorics (Adger & Svenonius, 2011). These objects can be combined in workspaces, phases or cycles (Frampton & Gutmann, 1999), which can be extended to form *non-local dependencies*. As we will discuss, these properties are guided by principles of minimal search (an optimal tree-search procedure, informed by notions from computer science; Aycock, 2020; Ke, 2019; Roberts, 2019) and least effort (Larson, 2015), akin to FEP formulations, fulfilling the imperatives of active inference to construct meaningful representations as efficiently as possible (Bouizegarene et al., 2020), directly contributing to surprise minimisation.

2.2 Compositionality

Recently, certain efficiency principles at the conceptual interface (where syntax interfaces with general conceptualization) have been proposed (Pietroski, 2018), such that the 'instructions' that language provides for the brain to build specific meanings are interpreted with notable simplicity. Leaving more technical details aside, this is ultimately achieved (in Pietroski's recent model) through M-join (e.g., F(_) + G(_) → F^G(_), which combines *monadic* concepts, like *red* + *boat*) and D-join (e.g., D(_,_) + M(_) → ∃[D(_,_)^M(_)], which merges *dyadic* concepts with monadic concepts, deriving the meaning of *X verb(ed) Y*). Hence, natural language permits limited dyadicity as a very minimal departure from the most elementary monadic combinatorial system. Human language is marginally more expressive (in its conceptual interpretations) than a first-order language (i.e., one set, and one embedding), but the interpretation system is the least complex needed to express dyadicity and permit *relations* between sets. As with models of syntax invoking a simple process of binary set-formation to derive recursion, by restricting the number of computational procedures able to generate semantic structures, this model restricts in highly predictable ways the possible range of outputs.

Consider also how, in neo-Davidsonian event semantics, conjunction is limited to predicates of certain semantic types (Pietroski, 2005; Schein, 1993). Certain semantic rules of composition, in (1b), have been claimed to arise directly from more elementary syntactic computations (Pietroski, 2018) which adhere to principles of efficient computation.

(1) a. Dutch shot Micah quickly.
b. ∃e[Agent(e, Dutch) & Theme(e, Micah) & shot(e) & quickly(e)]

In this connection, it has further been observed that language acts as an artificial context which helps "constrain what representations are recruited and what impact they have on reasoning and inference" (Lupyan & Clark, 2015, p. 283). Words themselves are "highly flexible (and metabolically cheap) sources of priors throughout the

Footnote 3 continued
for sensations. Perhaps the clearest example here is the carving of a sensory world into *what* and *where* (Ungerleider & Haxby 1994).

neural hierarchy" (Ibid) (for discussion of simplicity in semantic computation, see Al-Mutairi, 2014; Bošković & Messick, 2017; Collins, 2020; Gallego & Martin, 2018; González Escribano, 2005; Hauser et al., 2002; Hornstein & Pietroski, 2009). Overall, the entirety of the core language system (compositional syntax-semantics) appears to be shot through with criteria of efficiency that would inform any nascent generative model of linguistic syntax.

3 A Kolmogorov complexity estimate for narrow syntax

3.1 Economy

The notion of simplicity has been a methodological value which has guided linguistic inquiry for decades (Terzian & Corbalan 2021). Chomsky (2021b, p. 13) notes that "measuring simplicity is an essential task and is no simple matter". We aim in this section to elaborate a measure of syntactic complexity that connects directly to the principles that underwrite the FEP. We will not be concerned with typological, phonological or acquisitional notions of complexity, which form the bulk of current literature. Instead, we are interested in underlying representational issues that pertain to syntax-semantics. Even the most recent explorations of simplicity in language, such as the volume on simplicity in grammar learning in Katzir et al. (2021), focus on modelling minimum description length in phonology and morphology, or morphosyntactic complexity across distinct languages (Ehret et al., 2023), but not processes pertaining to the internal derivation of syntactic objects. Much of this work fruitfully explores complexity and simplicity across *languages* (Ehret et al., 2023), using measures such as second language acquisition difficulty and situational diversity (counting the range of communicative contexts a language can be used in), as opposed to the computational architecture of the language faculty itself. Even if we consider possible complexity metrics for 'syntax' as a global system, such as *degree of subordination* (how hierarchically deep a structure can be; e.g., McCarty et al., 2023), this only gets us part of the way; something much more generic and encompassing is needed.

A number of economy principles have been proposed in theoretical linguistics: the No Tampering Condition (Chomsky, 2008), Minimal Link Condition (Chomsky, 1995), Minimal Yield (Chomsky, 2021c), Extension Condition (Chomsky, 1995), Last Resort (Chomsky, 1995), Relativised Minimality (Rizzi, 1990, 2001), Inclusiveness Condition (Chomsky, 1995), Precedence Resolution Principle (Epstein et al., 1998), Scope Economy (Fox, 2000), Phase Impenetrability Condition (Chomsky, 2004), Full Interpretation (Freidin & Lasnik, 2011; Lohndal & Uriagereka, 2016), Global Economy Condition (Sternefeld, 1997), Feature Economy (van Gelderen, 2011), Accord Maximization Principle (Schütze, 1997), Input Generalisation (Holmberg & Roberts, 2014), Maximise Minimal Means (Biberauer, 2019a), Resource Restriction (Chomsky et al., 2019), and Equal Embedding (Murphy & Shim, 2020) (for further discussion, see Frampton & Gutmann, 1999; Fukui, 1996; Titov, 2020).

Although economy principles have long figured in models of phonology, morphology and the lexicon (e.g., the Elsewhere condition, underspecification), it is only relatively recently that theories of syntax have embraced economy not just as a heuristic guiding research, but more concretely as a constitutive principle of language design (Leivada & Murphy, 2021; Murphy, 2012, 2020a; Reuland, 2011; Sundaresan, 2020). These have been framed within a linguistic context, often invoking domain-specific notions (Wilder et al., 1997), despite a core part of the intended project of modern theoretical linguistics being to embed linguistic theory within principles general to cognition. Motivating language-specific computational generalizations by direct reference to the FEP may broaden the explanatory scope for the existence and prevalence of particular syntactic phenomena. Since linguists lack a general theory of computational efficiency for language (e.g., Gallego & Chomsky, 2020: "To be sure, we do not have a general theory of computational efficiency"), additional support with respect to grounding these concerns within a well-motivated framework for general organic behaviour will likely prove productive. Linguists have long speculated about how to model simplicity, but surprisingly few have done so rigorously. For example, Pearl (2022) speculates that "perhaps the knowledge of the tight relationship between syntax and meaning is some kind of simplicity bias that assumes maximal similarity between representational systems, unless shown otherwise". There are many promising paths to take here: minimising energy expenditure during language processing (Rahman & Kaykobad, 2005), shortening description length (Schmidhuber, 2015), reducing Kolmogorov complexity (Ming & Vitányi, 2008; Wallace & Dowe, 1999), and the degree of requisite belief updating. Relatedly, we might consult the principles of minimum redundancy and maximum efficiency in perception (Barlow, 1961, 1974, 2001; Wipf & Rao, 2007). We will provide a concrete exploration of one of these notions below (Kolmogorov complexity) in order to defend what we will term Turing–Chomsky Compression, with the immediate disclaimer that we acknowledge that other measures may in fact ultimately be more useful and well-motivated with respect to building FEP-syntax sympathies, and that we consider our survey to be purely preliminary.

A core fact about many natural language expressions is that they involve arrangements of nested constituents that enter into relations and dependencies of various kinds. How are syntactic operations compressed into determinate, unambiguous instructions to conceptual systems, and are there any general laws of organic design that can be inferred from the FEP that appear to constrain this process (and which successfully predict which objects *cannot* be constructed)?

Consider how under the No Tampering Condition the merging of two syntactic objects, X and Y, leaves X and Y unchanged. The set {X, Y} created by MERGE (Chomsky et al., 2019) cannot be broken and no new features can be added.[4] The original structure in (2a) can be modified by the merging of a new element, λ, to form (2b), adhering to the No Tampering Condition, while (2c) violates this condition

[4] MERGE has been defined as an operation on a workspace and its objects, formalized as follows (WS = workspace; P/Q = workspace objects such as linguistic features; X = additional elements):MERGE(P,Q,WS) = WS' = $\{\{P,Q\},X_1,\ldots,X_n\}$.Other linguistic frameworks assume some similar, basic structure-building operation, e.g., *Forward–Backward Application* in Combinatory Categorial Grammar (Steedman 2000), *Substitution* in Tree-Adjoining Grammar (Joshi & Schabes 1997). We put aside here some controversies about the relation between a MERGE-based syntax and set theory (see Adger Forthcoming, Gärtner 2021).

since the original structure (2a) is modified (Lohndal & Uriagereka, 2016); hence why adjuncts that merge 'downstairs' do not alter the structure of the object they adjoin to (adjuncts are not labeled; Bošković, 2015) (subscripts denote syntactic heads/labels, standard script denotes lexical items, where (2a) could represent a structure like 'the red boat').[5]

(2) a. $[_{\alpha}\ [\beta\ [_{\gamma}\ [\delta\ \varepsilon]]]]$
b. $[_{\alpha}\ \lambda\ [_{\alpha}\ [\beta\ [_{\gamma}\ [\delta\ \varepsilon]]]]]$
c. $[_{\alpha}\ [\beta\ [_{\gamma}\ [\delta\ [_{\varepsilon}\ \varepsilon\ \lambda]]]]]$

Further, it is more *economical* to expand a structure, as in (2b), than to backtrack and modify a structure that has already been built, as in (2c) (Lasnik & Lohndal, 2013). How can we more formally demonstrate these claims? We turn here to Kolmogorov complexity.

3.2 Compression

Kolmogorov complexity is a measure of the length of the shortest program that can reproduce a given pattern (Kolmogorov, 1965; Li & Vitányi, 2019). While measures of minimum description length and Kolmogorov complexity have been typically applied to linear, 'externalized' sequences, they can also be fruitfully applied to grammatical relations, permitting measurement of the inherent information content of an individual object or operation (Biberauer, 2019b, Grünwald 1996, 2007, Newmeyer & Preston, 2014). Sequence complexity is identified with richness of content (Mitchell, 2009), such that any given signal or sequence is regarded as complex if it is not possible to provide a compressed representation of it. While complexity differences across languages can be measured, for example, as a function of featural specifications on functional elements (Longobardi, 2017), here we are interested in the complexity of I-language derivations. Previous efforts have already connected the theory of program size to psychology by implementing a concrete language of thought with Turing-computable Kolmogorov complexity (Romano et al., 2013), which satisfied the requirements of (i) being simple enough so that the complexity of any binary sequence can be measured, and (ii) utilizing cognitively plausible operations like *printing* and *repeating*. In contrast, we aim to relate similar measures to syntactic economy criteria.

The concept of *syntactic complexity* remains underexplored in the literature relative to other measures of linguistic complexity (Shieber, 1985; Trudgill, 2011). While syntacticians have proposed economy principles, these all effectively boil down to efficient tree-search—without formalizing these intuitions or attempting to broach this topic with neighboring fields that might be able to provide highly valuable analytic tools. It is our contention that mathematical tools emerging from concerns of the FEP can help couch these verbal generalizations into a concrete model.

Syntactic complexity can be operationalized across a number of dimensions, such as online processing/parsing complexity (Hawkins, 2004), tree-search and node counts (Szmrecsányi, 2004), number of MERGE applications (Samo, 2021),

[5] Adjacent to minimalist syntax, Optimality Theory assumes that gratuitous adjunction to a maximal projection violates an economy condition (SPECLEFT) (Broekhuis & Vogel 2009; Grimshaw 2001). For syntactic economy conditions in Lexical Functional Grammar, see Dalrymple et al. (2015).

cyclic/derivational complexity (Trotzke & Zwart, 2014), internal representational complexity as opposed to derivational size (van Gelderen, 2011, 2021), entropy reduction (Hale, 2016), or stages of second-language development (Walkden & Breitbarth, 2019). Syntactic complexity can be framed as grammar-based (derivational), or user-based (parsing) (Newmeyer & Preston, 2014); here, will be elaborating on the former type. Crucially, sentence length does not always scale with syntactic complexity (Szmrecsányi, 2004), and instead an examination of the underlying operations is required. Although syntactic complexity is often thought of in derivational terms, removed and independent from surface realization, Kolmogorov complexity is relatively theory-neutral and can be applied indiscriminately to mental objects with any number of internal sequences, patterns, irregularities, and surface redundancies (Miestamo et al., 2008).

Why do we choose to focus here on such a generic, theory-neutral measure as opposed to a more domain-specific one? We stress that Kolmogorov complexity (and the related notion of minimal message length) relates directly to frameworks emerging from the FEP (Friston, 2019; Hinton & Zemel, 1994; Korbak, 2021), with the prediction for natural language syntax of reducing the complexity of hierarchical syntactic structures that are interpreted at conceptual interfaces being sympathetic to a corollary of the FEP that every computation comes with a concrete energetic cost (Jarzynski, 1997; Sengupta & Stemmler, 2014). As shown above (Eq. 2), variational free energy can be formulated as a trade-off between accuracy and complexity, whereby minimising complexity minimises variational free energy. Considering the topic of universal computation, as in Solomonoff induction (Solomonoff, 1964) (which is directly grounded in the minimization of Kolmogorov complexity), many formulations of variational free energy minimization explicitly invoke algorithmic complexity and the type of mathematical formulations underlying universal computation. Relating this more directly to our present concerns, the theme of message length reduction has been fruitfully applied to analyses of grammar acquisition in children. Rasin et al. (2021) show that minimum description length (closely related to Bayesian grammar induction) can provide the child with a criterion for comparing hypotheses about grammatical structures that may match basic intuitions across a number of cases. The *restrictiveness* generated by these complexity measures supplements the more general *simplicity* criterion of theoretical syntax; much as how the ‘subset principle’ (restrictiveness) supplemented the original evaluation metric (simplicity) (Berwick, 1985). Lambert et al. (2021) demonstrated that the computational simplicity of learning mechanisms appears to have a major impact on the types of patterns found in natural language, including for syntactic trees, and so it seems to us well motivated to turn to the issue of the underlying processes that guide the generation of these structures.

Other recent work has successfully used minimum description length in a domain much closer to our own concerns. Focusing on semantic properties of quantifiers (e.g., ‘some’, ‘most’) and noting that natural language quantifiers adhere to the properties of *monotonicity*, *quantity* and *conservativity* (Barwise & Cooper, 1981), van de Pol et al. (2021) generated a large collection of over 24,000 logically possible quantifiers and measured their complexity and whether they adhered to the three universal properties. They found that quantifiers that satisfied universal semantic properties were less complex and also exhibited a shorter minimal description length compared to quantifiers that did not satisfy the universals, pointing in intriguing directions towards efficiency biases in natural language semantics that appear to restrict the development of lexical meaning. Quantifiers that adhere to semantic universals are *simpler* than logically possible competitors that do not.

To briefly formalize our discussion of compression and complexity, given a Turing machine M, a program p and a string x, we can say that the Kolmogorov complexity of x relative to M is defined by the length of x. Formally, this can be expressed as follows (Eq. 4), where $|p|$ denotes the length of p and M is any given Turing machine:

$$K_M(x) \overset{\text{def}}{=} \min\{|p| : M(p) = x\} \cup \{\infty\} \tag{4}$$

This represents the length of the shortest program that prints the string x and then halts. Yet, as implied by Gödel's incompleteness theorem or Turing's halting theorem, we cannot compute the Kolmogorov complexity of an arbitrary string, given that it is impossible to test all possible algorithms smaller than the size of the string to be compressed, and given that we cannot know that the Turing machine will halt (Chaitin, 1995). We therefore used an estimate of approximate Kolmogorov complexity (given its fundamental non-computability) based on the Lempel–Ziv compression algorithm, which we applied to the labeling/search algorithm needed to derive each syntactic node in (2) and their subordinated terminal elements, investigating how 'diverse' the patterns are that are present in any given representation of a syntactic tree-search. In the service of replicability and for completeness, we used a current generative grammar labeling/search algorithm that checks tree-structure heads and terminal elements (Chomsky, 2013; Ke, 2019; Murphy & Shim, 2020) (see also f.n. 11). In this respect, Kolmogorov complexity is a more fine-grained measure of complexity than previous measures in theoretical syntax (e.g., node count across a c-commanded probe-goal path). While we acknowledge here and elsewhere (below) that our choice of algorithmic complexity is motivated by a similarity in spirit and mathematical heritage, we feel that this approach—given the relative novelty of connecting the FEP with theoretical linguistics—is suitably noncommittal with respect to which algorithmic complexity measure is ultimately going to show the most direct sympathy with syntactic derivational architecture, and below we will discuss some other possible measures.

Searching the structure from top to bottom, identifying each branching node and its elements (e.g., inputting [α λ α β γ δ ε]), we used a Lempel–Ziv implementation (Faul, 2021) of the classical Kolmogorov complexity algorithm (Kaspar & Schuster, 1987; Lempel & Ziv, 1976) to measure the number of unique sub-patterns when scanning the string of compiled nodes.[6] This Lempel–Ziv algorithm computes a Kolmogorov complexity estimate derived from a limited programming language that permits only copy and insertion in strings (Kaspar & Schuster, 1987).[7] The algorithm scans an n-digit sequence, $S = s_1 \cdot s_2 \cdot \ldots s_\mathrm{n}$, from left to right, and adds a new element to

[6] We therefore assume that labeling/search occurs top-to-bottom, and not bottom-up, due to the former yielding a more minimal search path with no 'backtracking', hence more in line with economy considerations.

[7] Kaspar and Schuster (1987) discovered that a readily calculable measure of Lempel–Ziv algorithmic complexity can, for simple cellular automata, separate pattern formation from the mere reduction of source

its memory each time it encounters a substring of consecutive digits not previously encountered. Our measure of Kolmogorov complexity takes as input a digital string and outputs a normalised measure of complexity (Urwin et al., 2017).

To connect these ideas with the FEP, we note that minimising free energy corresponds to minimising complexity, while maximising the accuracy afforded by internal representations[8] $r \in R$, of hidden states $s \in S$, given outcomes $o \in O$ (Eq. 2). In short, belief updating or making sense of any data implies the minimisation of complexity:

$$D_{KL}[Q(s)\|P(s)] \approx D_{KL}[P(s|o)\|P(s)] \tag{5}$$

When choosing how to sample data, the expected complexity becomes the intrinsic value or expected information gain (in expected free energy):

$$\mathbb{E}[lnP(s|o,\pi) - lnP(s|\pi)] = I(S,O|\pi) = \mathbb{E}_{P(o|\pi)}[D_{KL}[P(s|o,\pi)\|P(s|\pi)]] \tag{6}$$

This is just the mutual information between (unobservable) hidden states generating (observable) outcomes, under a particular choice or policy.

Importantly, variational free energy and formulations of artificial general intelligence pertaining to universal computation both share a mathematical legacy. This is rooted in the relationship between the complexity term in variational free energy and algorithmic complexity (Hinton & Zemel, 1994; Wallace & Dowe, 1999), described in terms of information length and total variation distance. As such, relating syntactic operations to algorithmic compression maximisation feeds directly into assumptions from the FEP (Schmidhuber, 2010).

Lempel–Ziv complexity is a measure of algorithmic complexity which, under the law of large numbers, plays the same role as the complexity part of log model evidence or marginal likelihood. Interestingly, minimising algorithmic complexity underwrites universal computation, speaking to a deep link between compression, efficiency and optimality in message passing and information processing. This is why we suggest here that although we are using Lempel–Ziv complexity, we suspect that many other, potentially more suitable measures of compressibility might be used in future to relate the FEP to syntax.

With this background, we now return to the structures in (2b) (licensed) and (2c) (unlicensed). Inputting the labeled nodes (subscript elements) and terminal elements (regular script) across both structures into the Lempel–Ziv compression algorithm (Faul, 2021) left-to-right, the licensed representation in (2b) exhibits a normalized Kolmogorov complexity of 1.88, while the unlicensed representation in (2c) exhibits a complexity of 1.99. Crucially, while both (2b) and (2c) exhibit the same node-count complexity and depth (i.e., bracket count), they can be operationally distinguished by their Kolmogorov complexity, in compliance with what the FEP would demand.

Footnote 7 continued
entropy, with different types of automata being able to be distinguished. We also note that Lempel–Ziv complexity does not simply measure the number of elements in a sequence, but also factors in pattern irregularities. As such, it is not the case that, by definition, a syntactic process with n steps will be trivially more Lempel–Ziv complex than a syntactic process with n-1 steps.

[8] That parameterise posterior beliefs $Q(s) \triangleq Q_r(s)$.

The increased compression rate for (2b) indicates lower information content, hence lower Kolmogorov complexity (Juola, 2008), and so the representation adheres to the priority to minimise computational load.

3.3 Relativised minimality

A further observation pertaining to economy in the literature concerns Relativised Minimality (Rizzi, 1990, 1991): Given a configuration, [X … Z … Y], "a local relation cannot connect X and Y if Z intervenes, and Z fully matches the specification of X and Y in terms of the relevant features" (Starke, 2001). In other words, if X and Y attempt to establish a syntactic relation, but some element, Z, can provide a superset of X's particular features (i.e., X's features plus additional features), this blocks such a relation. In (3a), *which game* provides a superset of the features hosted by *how*, resulting in unacceptability. The equivalent does not obtain in (3b), and so a relationship between both copies of *which game* can be established (strikethroughs denote originally merged positions).

(3) a. *[[How$_{[+Q]}$] [C$_{[+Q]}$ [do you wonder [[which game$_{[+Q,\ +N]}$] [C$_{[+Q]}$ [PRO to play ~~how~~$_{[+Q]}$]]]]]].
b. [[Which game$_{[+Q,+N]}$] [C$_{[+Q]}$ [do you wonder [[how$_{+Q}$] [C$_{+Q}$ [PRO to play ~~which game~~$_{[+Q,+N]}$]]]]]]

Relativised Minimality emerges directly from minimal search (Aycock, 2020): Consider how when searching for matching features in (3b) the search procedure would skip *how* but find the original copy of *which game*.

We note that the notion of movement 'distance' here is relativised to the specific units across the path. In order to reach a more fundamental analysis we may need some means of understanding what distance actually reduces to. These avenues of current research may lend themselves quite readily to explorations directed by notions of complexity and compression. We speculate here that this may relate to the compressibility of movement paths, and more systematic investigations will be needed to address this.

3.4 Resource restriction

The principle of Resource Restriction (or 'Restrict Computational Resources', RCR; Chomsky, 2019; Chomsky et al., 2019) states that when the combinatorial operation MERGE maps workspace n to workspace $n + 1$, the number of computationally accessible elements (syntactic objects) can only increase by one (Huybregts, 2019; Komachi et al., 2019). This can account for a peculiar property of natural language recursion that separates it from other forms of recursion (e.g., propositional calculus, proof theory): natural language MERGE involves a recursive mapping of workspaces that removes previously manipulated objects (Chomsky, 2021c). Hence, Resource Restriction renders natural language derivations strictly Markovian: The present stage is independent of what was generated earlier, unlike standard recursion. MERGE itself exhibits the formal characteristics of a finite-state rewrite rule (Trotzke & Zwart, 2014,

p. 145), exhibiting minimal computational complexity, with MERGE being distinct from the ultimate grammatical constructions later derived from its cyclic application. Even though this is in line with traditional assumptions from generative grammar that there is something unique to human syntax (which we concur with), we wish to stress that the formalization of this property of recursion does not necessitate complete isolation from domain-general approaches in the cognitive sciences; i.e., there are means to ground and explain this property through models emerging from the FEP.

A topic of recent discussion concerns how we can define the ‘size’ of a workspace. Fong et al. (2019) suggest that the size of a syntactic workspace should be considered to be the number of accessible terms plus the number of syntactic objects. This proposal to constrain syntactic combinatorics can account for why the applications of certain types of MERGE are ungrammatical (Fong et al., 2019), providing a genuine explanation for language design. We again return to the theme of every computational procedure delivering a concrete cost, as under the FEP.

Principles such as Resource Restriction and other economy considerations are essential once we consider that a workspace with two elements with a simple MERGE operation can generate excessive levels of combinatoriality. Within 8 MERGE steps from two elements, around 8 million distinct sets can be formed (Fong & Ginsburg, 2018). Older definitions of basic syntactic computations did not “worry about the fact that it’s an organic creature carrying out the activities”, as Chomsky (2020) notes. Many aspects of these theories exhibited, to borrow a phrase from Quine (1995, p. 5), an “excess of notation over subject matter”. Even many current models of syntax have ignored questions of cognitive, implementational plausibility (e.g., Chomsky, 2013; Citko & Gračanin-Yuksek, 2021; Collins, 2017, Epstein et al., 2021). Computational tractability (van Rooj & Baggio, 2021) is a powerful constraint in this respect (e.g., implementable in polynomial time), and given that minimizing the model complexity term (in formulations of free energy) entails reducing computational cost, this efficiency constraint is also implicitly present in the FEP.

3.5 Interim conclusion

We have considered how the FEP can in principle provide a novel explanation for the prevalence of efficiency-encoded syntactic structures. To further stress this point, consider Dasgupta and Gershman’s (2021) assessment that mental arithmetic, mental imagery, planning, and probabilistic inference all share a common resource: memory that enables efficient computation. Other domains exhibiting computational efficiency include concept learning (Feldman, 2003), causal reasoning (Lombrozo, 2016) and sensorimotor learning (Genewein & Braun, 2014). As Piantadosi (2021) reviews, human learners prefer to induce hypotheses that have a shorter description length in logic (Goodman et al., 2008), with simplicity preferences possibly being “a governing principle of cognitive systems” (Piantadosi, 2021, p. 15; see Chater & Vitányi, 2003). Although our arguments have been almost exclusively conceptual, we believe that more extensive computational modelling should seek to compare the dynamics of MERGE-based workspaces via compressibility constraints.

We will now turn to the most commonly explored syntactic processes claimed to arise from economy considerations: syntactic movement and minimal search. Further examples will be used to motivate what we term the principle of Turing–Chomsky Compression, through which stages of syntactic derivations are evaluated based on the algorithmic compressibility of some feature of the computation, such as the movement path of an object, or the procedure of nodal labeling/search—which can be unified based on how they manipulate the syntactic workspace. Turing–Chomsky Compression provides a concrete architectural principle for language design, which is crucially sympathetic to a number of compressibility criteria beyond the types entertained here; we therefore conclude with a number of promising directions to refine this new approach to syntax.

4 Minimising free-energy, maximising interpretability

As has long been recognised, the syntactic categories of words are not tagged acoustically, and yet sentential meaning is inferred from syntactic categorization (Adger, 2019). Sentences are often ambiguous between distinct syntactic structures. For instance, below we can interpret Jim Carrey as starring in the movie (4a), or sitting next to us (4b).

(4) a. [$_{\mathrm{TP}}$ [$_{\mathrm{NP}}$We] [$_{\mathrm{VP}}$watched [$_{\mathrm{NP}}$a [$_{\mathrm{N}}$movie [$_{\mathrm{PP}}$with [$_{\mathrm{NP}}$Jim Carrey]]]]]].
b. [$_{\mathrm{TP}}$ [$_{\mathrm{NP}}$We] [$_{\mathrm{VP}}$[$_{\mathrm{VP}}$watched [$_{\mathrm{NP}}$a movie]] [$_{\mathrm{PP}}$with [$_{\mathrm{NP}}$Jim Carrey]]]].

Linear distance (i.e., the number of intervening elements between dependents in a sentence) can be contrasted with structural distance (the number of hierarchical nodes intervening), and only the latter is a significant predictor of reading times in an eye-tracking corpus (Baumann, 2014). Violations of hierarchical sentence rules results in slower reading times (Kush et al., 2015), and expectations of word category based on hierarchical grammars also predicts reading times (Boston et al., 2011).

The apparent use of hierarchical structure to *limit* interpretation adheres to a core tenet of the FEP, whereby interpretive processes that yield the lowest possible amount of complexity (and thereby computational cost) can mostly (perhaps entirely; Hinzen, 2006) be derived directly from what the syntactic component produces. This notion is closely related to the imperatives for structure learning (Tervo et al., 2016)—or Bayesian model reduction—in optimising the structural (syntax) of generative models based purely on complexity (pertaining to model parameters); see Friston et al. (2017b) for an example simulating active inference and insight in rule learning.

While sensorimotor systems naturally impose linear order, linguistic expressions are complex *n*-dimensional objects with hierarchical relations (Gärtner & Sauerland, 2007; Grohmann, 2007; Kosta et al., 2014; Murphy & Benítez-Burraco, 2018; Murphy, 2020b, 2024). The following sections provide concrete demonstrations of these design principles in action in order to motivate an architectural framework for language emerging from the FEP. We also provide suggestions for how to explore further sympathies between the FEP and minimalist syntax.

4.1 Structural distance

Consider the sentence in (5).

(5) Routinely, poems that rhyme evaporate.

In (5), ‘routinely’ exclusively modifies ‘evaporate’. The matrix predicate ‘evaporate’ is closer in terms of *structural distance* to ‘routinely’ than to ‘rhyme’, since the relative clause embeds ‘rhyme’ more deeply (minimal search is partly “defined by least embedding”; Chomsky, 2004, p. 109).[9] Language computes over structural distance, not linear distance (Berwick et al., 2011, 2013; Friederici et al., 2017; Martin et al., 2020).

This can also be shown with simple interrogative structures. Consider the sentence in (6a) and its syntactic representation in (6b), where the verb in the relative clause (‘rhyme’) is more deeply embedded than ‘evaporate’.

(6) a. Do poems that rhyme evaporate?
b. [CP[C Do][TP[DP[DP poems][CP[C that][TP rhyme]]][T’[T][V evaporate]]]].

We can compute the complexity of both nodal search and Kolmogorov complexity, contrasting the grammatical association between ‘Do’ and ‘evaporate’, and the ungrammatical association between ‘Do’ and ‘rhyme’. When the [+ Q] feature on C searches for a goal, it needs to search down three node steps (from CP to V) to get to the grammatical option, but needs to search down four node steps (from CP to embedded TP) to construct the ungrammatical option. Since we are concerned with analyzing a small but representative number of syntactic derivational processes, this analysis differs from the approach to the structures in (2), which do not involve any labeling procedure. This time, we enumerated the search steps across nodes, replacing specific nodal categories with symbols interpretable to the Lempel–Ziv compression algorithm (Faul, 2021), since this is what the syntactic search algorithm is claimed to monitor. The Lempel–Ziv complexity of the sequence of steps enumerated from the C-V labeling/search algorithm is 1.72. For the embedded C-TP search, it increases to 2.01.

While one might invoke purely semantic constraints on polar interrogatives and other forms of question-formation (Bouchard, 2021) to derive the kinds of acceptability contrasts we have discussed, we see no way to ground these observations in concerns of computability and complexity, and no way to quantify or formalize these notions.

[9] We refer the reader to Ke (2019, p. 44) and Aycock (2020, pp. 3–6) for a detailed discussion of minimal search, which can be formally defined below, from Aycock (2020), adopting an Iterative Deepening Depth-First Search approach (Korf 1985); where MS = minimal search, SA = search algorithm, SD = search domain (where SA operates), ST = search target:

(1) MS = ⟨SA, SD, ST⟩
(2) SA:

a. Given ST and SD, match against every head member of SD to find ST [initial depth-limit of SD = 1; search depth-first].b. If ST is found, return the head(s) bearing ST and go to d. Otherwise, go to c.c. Increase the depth-limit of SD by 1 level; return to a.d. Terminate Search.

4.2 Ignoring other people: question formation via economy

As recent literature has explored, whenever there is a conflict between principles of computational efficiency and principles of communicative clarity, the former seems to be prioritized (Asoulin, 2016; Murphy, 2020a). For instance, consider (7).

(7) You persuaded Saul to sell his car.

The individual (‘Saul’) and the object (‘car’) can be questioned, but questioning the more deeply embedded object forces the speaker to produce a more complex circumlocution (‘[]’ denotes the originally merged position of the *wh*-expression).

(8) a. *[What] did you persuade who to sell []?
b. [Who] did you persuade [] to sell what?

The structures in (8) involve the same words and interpretations, yet the more computationally costly process of searching for—and then moving—the more deeply embedded element cannot be licensed, despite the potential benefits of communicative flexibility. Interestingly, one cannot feasibly posit parsing-related factors to derive some independent complexity measure to explain this contrast (e.g., Newmeyer, 2007), given the same number of words and same semantic interpretations (i.e., *give me the Agent and Object of the event*). Experimental work has supported the prevalence of these grammaticality intuitions (Clifton et al., 2006).[10]

The syntactic structures for both (8a) and (8b) are represented in (9) (where < DP > represents the movement path). With respect to tree-search depth, (9a) involves searching down 11 nodes, while (9b) involves searching down 9 nodes. To expand our survey of syntactic processes beyond labeling/search paths, we focused here on the postulated path of syntactic object movement across the structure. The movement path was represented with each site being attributed a symbol fed to the compression algorithm, in keeping with a more general approach to annotating movement paths (Adger, 2003). Enumerating the movement path from the initially merged root, to intermediate landing sites, to the terminal landing site in Spec-CP, the Lempel–Ziv complexity of movement for (9a) is 2.15. For (9b), path complexity is 1.5.

(9) a. [$_{\text{CP}}$ [$_{\text{DP}}$ what] [$_{\text{C'}}$ [$_{\text{C}}$ did] [$_{\text{TP}}$ [$_{\text{DP}}$ you] [$_{\text{T'}}$ [$_{\text{T}}$ [*pres*] [$_{\text{VP}}$ [$_{\text{<DP>}}$] [$_{\text{V'}}$ [$_{\text{V}}$ persuade] [$_{\text{CP}}$ [C’ [C Ø] [TP [<DP>] [T’ [T] [VP [DP who] [PP [P to] [VP [V’ [V sell] [<DP>]]]]]]]]]]]]]]]]].
b. [CP [DP who] [C’ [C did] [TP [DP you] [T’ [T [pres] [VP [<DP>] [V’ [V persuade] [CP [C’ [C Ø] [TP [<DP>] [T’ [T] [VP [<DP>] [PP [P to] [VP [V’ [V sell] [DP what]]]]]]]]]]]]]]]]

A further empirical reason to assume that this economy condition is a general property of language comes from the following data of Bulgarian multiple *wh*-fronting (Bošković & Messick, 2017; see also Dayal, 2017). The *wh*-phrase highest prior to movement (the subject in (10) and the indirect object in (11)) needs to be first in the

[10] We also highlight here the generalisation, discussed extensively in Jackendoff and Wittenberg (2014), that simpler syntactic structures typically lead to a greater reliance on pragmatics for successful communication, whereas larger sentences lead to more of an interpretive burden being placed on syntactic principles instead of conversational context.

linear order of the sentence, such that the structurally highest *wh*-phrase moves first, and the second *wh*-phrase either right-adjoins to the first *wh*-phrase, or moves to a structurally lower Spec-CP position.

(10) a. *Koj e vidjal kogo?
who is seen whom
b. Koj kogo e vidjal?
"Who saw whom?"
(11) a. Kogo kakvo e pital Ivan?
whom what is asked Ivan
"Whom did Ivan ask what?"
b. *Kakvo kogo e pital Ivan?

Thus far, this suffices to show that the *wh*-element easiest to search for is selected for movement. However, does syntactic economy simply rule out all but one option? Crucially, Bošković and Messick (2017) show that when multiple options of equal tree-geometric complexity are available, they are *both* licensed as grammatical. Consider constructions with three *wh*-phrases. We can assume that whichever *wh*-element moves to the structurally highest position (Spec-CP) satisfies the featural requirement of interrogative C to have a filled Spec-CP position. After this structurally highest element moves to Spec-CP, we can further assume that the remaining *wh*-elements then move to Spec-CP to satisfy their own featural 'Focus'-based requirements. At this point, whichever order the remaining *wh*-elements move in, the requirements are satisfied through movements of identical length (i.e., both cross the same number of nodes, and hence generate the same sequence of derivational steps, and therefore the same Lempel–Ziv complexity). As such, this predicts that the remaining two *wh*-elements can move in any order after the initial *wh*-movement of the subject. This prediction is borne out: the subject ('koj') is moved first in both constructions below, but then either of the remaining *wh*-elements can move in any order.

(12) a. Koj kogo kakvo e pital?
who whom what is asked
"Who asked whom what?"
b. Koj kakvo kogo e pital?

4.3 Labeling

As a more stringent test, can Lempel–Ziv complexity shed light on cases in which the ungrammatical derivation has *less* structural tree-geometric complexity than the grammatical derivation? Consider the following case from Murphy and Shim (2020, p. 204). Putting ancillary technical details aside (see Mizuguchi, 2019), (13a) is claimed to be ungrammatical because one final necessary operation on the syntactic workspace has not been carried out; namely, merging 'the students' to the structure marked by γ. For expository purposes, we provide a schematic representation to demonstrate the relevant movement path (the path of 'the student' is marked by *t*).

(13) a. *[$_\gamma$ Seems to be likely [$_\alpha$ the student [to [*t* understand the theory]]]].

b. [$_\delta$ The student [$_\gamma$ seems to be likely [$_\alpha$ *t* [to [*t* understand the theory]]]]]

The explanations from within syntactic theory as to why (13b) is grammatical concern successful feature valuation and the minimal search of copies via the labeling algorithm. However, this process might also be linked to more efficient compression rates of syntactic labels at the interpretive systems. We can enumerate each labeled node left-to-right marking the phrase boundaries separating each embedded object that pertain to the grammaticality contrast (e.g., V-D-P-V). Computing the Lempel–Ziv complexity of each successive phrase label in these structures, (13a) exhibits a complexity of 1.86, while (13b) exhibits a complexity of 1.66, despite (13b) being a more complex structure from the perspectives of node count and element count. As such, both minimal search of syntactic labels and algorithmic compression rates may be playing separate but interacting roles in determining how the interpretive systems access objects generated by syntax.

4.4 Turing–Chomsky Compression

The brief number of cases we have derived syntactic economy principles from, using a Lempel–Ziv estimate of Kolmogorov complexity, can be used to motivate the following language design principle that directly relates the FEP to syntactic structure building:

> ***Turing–Chomsky Compression*** An operation (*M*) on an accessible object (O_1) in a syntactic workspace (W_p) minimizes variational free energy if structures from the resulting workspace (W_q) are compressed to a lower Kolmogorov complexity than if *M* had accessed O_2 in W_p.

This is principally named after specifications over *what* (Chomsky) is compressed and *how* (Turing) such compression can be achieved (Chomsky, 2021c; Turing, 1950). The interaction between Turing–Chomsky Compression (TCC) and more domain-specific subcategorization requirements emerging from lexico-semantic features, and formal syntactic features, is a promising topic for future research. This will require a more accelerated survey of cognitive models of semantics that emerge from the FEP, including a more extensive and formal assessment of how to generate specific active components (under active inference) for lexico-semantic processing, with this stage coming into greater relevance after the initial generation of an abstract hierarchical structure that feeds, for example, event semantics. For now, we have shown across a small but representative number of syntactic processes that derivations minimising algorithmic complexity are licensed over those that result in structures and derivational paths that are less compressible.

We stress here that TCC is a concrete architectural proposal for language design, and is the type of principled settlement that could be established between mathematical models of compressibility that share a lineage with the FEP on the one hand, and models of minimalist syntax on the other. With this in mind, we now discuss some prospects for future research that could address these issues more systematically.

5 Future work

> Language and thought, in anything remotely like the human sense, might indeed turn out to be a brief and rare spark in the universe, one that we may soon extinguish. We are seeking to understand what may be a true marvel of the cosmos.
>
> Chomsky (2021c, p. 4)

We have arrived at a number of suggestive explanations for how language implements the construction of hierarchical syntactic objects: to minimise the computational burden of reading syntactic instructions at conceptual systems; to minimise uncertainty about the causes of sensory data; and to adhere to a least effort natural law (i.e., variational principle of least action) when composing sets of linguistic features for interpretation, planning and prediction. We have shown that measuring a Kolmogorov complexity estimate of syntactic representations and movement paths can align with acceptability judgments. This was used to motivate a possible principle of language design that could emerge from this research direction, Turing–Chomsky Compression (TCC). Our use of Lempel–Ziv complexity presents a more explicit measure than previous accounts. For instance, consider Sternefeld's (1997) *Global Economy Condition*, which states that, given two derivations of a syntactic structure (D1, D2), D1 is preferred if D1 fares better than D2 with respect to some metrical measure M (namely, number of derivational steps). This basic 'step counting' measure (as with tree-search depth) seems to be in line with grammaticality predictions emerging from the more general complexity measure provided by Lempel–Ziv complexity. Yet, algorithmic complexity also benefits from being applicable across a range of other domains in syntax where nodal count does *not* differ between competing structures, and is also related to formulations of variational free energy minimization. Ultimately, this has the advantage of generating quantitative predictions for syntactic computation based on general principles that apply more broadly.

Following neighbouring research in the active inference framework (Da Costa et al., 2021), one could feasibly view our research programme as comparing the information length of belief updating between distinct syntactic derivations and theories. We view our proposals as being, in principle, concordant with the view that neural representations in organic agents evolve by approximating steepest descent in information space towards the point of optimal inference (Da Costa et al., 2021). Future work could explore the utility of minimum description length (van de Pol et al., 2021) and Gell-Mann/Lloyd 'effective complexity'. In contrast to Kolmogorov complexity, which measures the description length of a whole object, effective complexity measures the description length of regularities (structured patterns) within an object (Gell-Mann & Lloyd, 1996), which may speak to properties of cyclic, phasal computation in natural language.

Recent work has provided evidence for a mental compression algorithm in humans (termed the Language of Thought chunking algorithm) responsible for parsing very basic, binary sequences, providing evidence that human sequence coding involves a form of internal compression using language-like nested structures (Planton et al., 2021). Dehaene et al. (2022) extend this project to auditory sequences and geometrical shapes. We have effectively extended these ideas further into the domain of natural language syntax, suggesting some common capacity for symbolic recursion across cognitive systems being constrained by compressibility.

Some linguists might object to our complexity measure in the following way: Why should syntax be organized so as to produce structures that minimise Kolmogorov complexity, and why should the semantic component of language aim to read off structures that are of a corresponding level of complexity? We note here that the core 'phase'/non-phase pattern of syntactic derivations (e.g., {C {T *v* {V D/*n* {N}}}}; Richards, 2011, Uriagereka, 2012) optimizes compression rate (effectively, 010101), and since phase construction constitutes the major determining period when syntactic workspaces are accessed by the conceptual systems, we see our proposal as aligning closely with existing—if only implicit—assumptions.

Plainly, there are many issues with the framework we have outlined here that need to be further unpacked and clarified. Our proposals concerning compression of structures accessed from the syntactic workspace via TCC have been discussed in the context of a cursory overview of the mathematical lineage shared between formulations pertaining to the FEP and theories of universal computation. This suited our current expository purposes, with our proposals being buttressed by conceptual overviews of the FEP and syntactic economy, but future work should more carefully align models emerging from the FEP with TCC. Although we made our assumptions about syntactic complexity based on whether or not our measure can be formally grounded within the FEP, we note that we have effectively equated complexity with compressibility. As such, we acknowledge that there may be a number of other fruitful directions to measure complexity in ways that are sympathetic to the FEP (e.g., the "complexity equals change" framework; Aksentijevic & Gibson, 2012).

Lastly, we acknowledge that our choice of complexity metric (Lempel–Ziv complexity) could well be argued to be sub-optimal, or even inappropriate for our focus on derivational syntax, and we hope to explore different varieties of compression algorithms that share a mathematical lineage with the FEP moving forward. This does not detract from our main conceptual arguments, which have constituted the bulk of our discussion, and nor does it preclude TCC being subject to further modification pending an expansion of compression algorithms tested, but we wish to note here that Lempel–Ziv is an optimally appropriate estimator for Kolmogorov complexity given long sequences produced by an independent and identically distributed ('iid') source; when Lempel–Ziv is applied to very short sequences, it becomes more sensitive to the structure of the compression algorithm. Concurrently, our sequences used to compute Lempel–Ziv complexity are produced by a non-iid (Markovian) source, and so would become increasingly redundant as string length increases. Hence, it may be the case that our choice of compression algorithm is non-optimal for both very short and longer sequences, and future work should seek to contrast multiple minimal description and compression algorithms jointly, potentially modifying the architecture of TCC.

6 Conclusions

We have reviewed how the FEP is an expression of the principle of least action, which is additionally a principle implemented in models of theoretical syntax. An intuition from 70 years ago—that the extent to which a model of grammar is simple seems to determine its explanatory power—echoes in the modern framing of syntactic computation as a system of economy: "[T]he motives behind the demand for economy are in many ways the same as those behind the demand that there be a system at all" (Chomsky, 1951; see also Goodman, 1951). Generative linguistics has long appealed to economy considerations (e.g., the evaluation metric in Chomsky & Halle, 1968). Meanwhile, the FEP has produced formal, simulation-supported models of complex cognitive mechanisms such as action, perception, learning, attention and communication, while theories of syntax embracing computational efficiency have led to empirically successful outcomes, explaining grammaticality intuitions (Adger, 2003; Martin et al., 2020; Sprouse, 2011; Sprouse & Almeida, 2017), certain poverty of stimulus issues (Berwick et al., 2011; Crain et al., 2017; Culbertson et al., 2012; Wexler, 2003; Yang et al., 2017) and the pervasive organizational role that hierarchy has in language (Friederici et al., 2017; Grimaldi, 2012; McCarty et al., 2023) and the seemingly unique proclivity humans have to parse sequences into tree-structures.

We find seeds for these ideas in the foundational principles of universal computation, where, as we have noted, free energy is often discussed in terms of minimum description or message lengths (MacKay, 2003; Schmidhuber, 2010). Relatedly, findings from dependency length minimization (DLM) research suggest that, during online parsing, comprehenders seek to minimize the total length of dependencies in a given structure since this reduces working memory load (Gibson et al., 2019).

A core objective of current theories of syntax is to explain *why* language design is the way it is (Adger, 2019; Narita, 2014), and we have suggested that the FEP can contribute to this goal. The more efficiently a language user can internally construct meaningful, hierarchically organized syntactic structures, the more readily they can use these structures to contribute to the planning and organization of action, reflection and consolidation of experience, exogenous and endogenous monitoring, imagination of possible states, adaptive environmental sampling, and the consideration of personal responsibilities. We used a brief number of examples to demonstrate proof of concept for how compression algorithms, such as a Kolmogorov complexity estimate, can provide principled insight into efficiency concerns alongside more traditional economy criteria such as node count and tree-search depth.

Moving beyond, we note that our measures of Kolmogorov complexity in the operations of natural language syntax are exploiting a highly general, theory-neutral measure of complexity, and that these and other related measures serve to index some underlying, independent process of organic computation, which remains elusive in its formal character and neural basis.

More broadly, what the FEP can offer theoretical linguistics is proof of principle: a foundational grounding and means of additional motivation for investigating language in terms of efficient computation. The FEP is fundamentally a normative model (Allen 2018) which can aid the generation of implementational models and can place constraints on feasibility. Further simulation and modeling work is required to push these ideas further for natural language and its various sub-systems, and we envisage that this type of future work will provide fruitful insights into natural language syntax.

Funding Open access funding provided by University College London. KF received funding from the Wellcome Centre for Human Neuroimaging (Ref: 205103/Z/16/Z) and a Canada-UK Artificial Intelligence Initiative (Ref: ES/T01279X/1). EH received funding from RNID (PA25_Holmes).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Adger, D. (2003). *Core syntax: A minimalist approach*. Oxford University Press.

Adger, D. (2019). *Language unlimited: The science behind our most creative power*. Oxford University Press.

Adger, D. (Forthcoming). *Mereological syntax: Phrase structure, cyclicity, and islands*. MIT.

Adger, D., & Svenonius, P. (2011). Features in minimalist syntax. In C. Boeckx (Ed.), *The Oxford handbook of linguistic minimalism* (pp. 27–51). Oxford University Press.

Aksentijevic, A., & Gibson, K. (2012). Complexity equals change. *Cognitive Systems Research, 15–16*, 1–16.

Al-Mutairi, F. R. (2014). *The minimalist program: The nature and plausibility of Chomsky's biolinguistics*. Cambridge University Press.

Allen, M. (2018). The foundation: mechanism, prediction, and falsication in Bayesian enactivism. Comment on Answering Schrödinger's question: a free-energy formulation, by Maxwell James Desormeau Ramsteal et al. *Physics of Life Reviews, 24*, 17–20.

Andrews, M. (2021). The math is not the territory: Navigating the free energy principle. *Biology & Philosophy, 36*, 30.

Asoulin, E. (2016). Language as an instrument of thought. *Glossa: A Journal of General Linguistics, 1*(1), 46.

Aycock, S. (2020). A third-factor account of locality: Explaining impenetrability and intervention effects with minimal search. *Cambridge Occasional Papers in Linguistics, 12*(1), 1–30.

Barandiaran, X. E., & Chemero, A. (2009). Animats in the modeling ecosystem. *Adaptive Behavior, 17*(4), 287–292.

Barlow, H. (1961). Possible principles underlying the transformations of sensory messages. In W. Rosenblith (Ed.), *Sensory communication* (pp. 217–234). MIT.

Barlow, H. B. (1974). Inductive inference, coding, perception, and language. *Perception, 3*, 123–134.

Barlow, H. (2001). Redundancy reduction revisited. *Computation and Neural Systems, 12*, 241–253.

Barwise, J., & Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy, 4*(2), 159–219.

Bastos, A. M., Martin Usrey, W., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron, 76*, 695–711.

Baumann, P. (2014). Dependencies and hierarchical structure in sentence processing. *Proceedings of CogSci, 2014*, 152–157.

Berwick, R. C. (1985). *The acquisition of syntactic knowledge*. MIT.

Berwick, R. C., Friederici, A. D., Chomsky, N., & Bolhius, J. J. (2013). Evolution, brain, and the nature of language. *Trends in Cognitive Sciences, 17*(2), 89–98.

Berwick, R. C., Pietroski, P., Yankama, B., & Chomsky, N. (2011). Poverty of the stimulus revisited. *Cognitive Science, 35*(7), 1207–1242.

Berwick, R. C., & Stabler, E. P. (Eds.). (2019). *Minimalist parsing*. Oxford University Press.

Biberauer, T. (2019a). Factors 2 and 3: towards a principled approach. *Catalan Journal of Linguistics Special Issue*, 45–88.

Biberauer, T. (2019b). Some thoughts on the complexity of syntactic complexity. *Theoretical Linguistics, 45*(3–4), 259–274.

Bošković, Ž. (2015). From the complex NP constraint to everything: On deep extractions across categories. *The Linguistic Review, 32*, 603–669.

Bošković, Ž. (2016). On the timing of labeling: Deducing Comp-trace effects, the Subject Condition, the Adjunct Condition, and tucking in from labeling. *The Linguistic Review, 33*, 17–66.

Bošković, Ž. (2021). *Merge, move, and contextuality of syntax: The role of labeling, successive-cyclicity, and EPP effects*. University of Connecticut.

Bošković, Ž, & Lasnik, H. (Eds.). (2007). *Minimalist syntax: The essential readings*. Blackwell.

Bošković, Ž, & Messick, T. (2017). Derivational economy in syntax and semantics. In M. Aronoff (Ed.), *Oxford research encyclopedia of linguistics.* Oxford University Press.

Boston, M., Hale, J., Vasishth, S., & Kliegl, R. (2011). Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes, 26*, 301–349.

Bouchard, D. (2021). Three conceptions of nativism and the faculty of language. *Languages Sciences, 85*, 101384.

Bouizegarene, N., Ramstead, M., Constant, A., Friston, K., & Kirmayer, L. (2020). Narrative as active inference. *PsyArXiv*. https://doi.org/10.31234/osf.io/47ub6

Bourguignon, M., Molinaro, N., Lizarazu, M., Taulu, S., Jousmäki, V., Lallier, M., Carreiras, M., & Tiège, X. D. (2020). Neocortical activity tracks the hierarchical linguistic structures of self-produced speech during reading aloud. *NeuroImage, 216*, 116788.

Broekhuis, H., & Vogel, R. (Eds.). (2009). *Optimality theory and minimalism: Interface theories*. Universitätsverlag Potsdam.

Brown, H., Adams, R. A., Parees, I., Edwards, M., & Friston, K. (2013). Active inference, sensory attenuation and illusions. *Cognitive Processing, 14*(4), 411–427.

Cardinaletti, A., & Starke, M. (1999). The typology of structural deficiency: A case study of the three classes of pronouns. In H. van Riemsdijk (Ed.), *Clitics in the languages of Europe* (pp. 145–233). Mouton de Gruyter.

Chaitin, G. J. (1995). Randomness in arithmetic and the decline and fall of reductionism in pure mathematics. In J. Cornwell (Ed.), *Nature's imagination* (pp. 27–44). Oxford University Press.

Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences, 7*(1), 19–22.

Chomsky N. (1949). Morphophonemics of modern Hebrew. Undergraduate Honors Thesis, University of Pennsylvania, Philadelphia.

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory, 2*, 113–124.

Chomsky, N. (1959). On certain formal properties of grammars. *Information and Control, 2*, 137–167.

Chomsky N. (1951/1979). *Morphophonemics of modern Hebrew*. Garland.

Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. Praeger.

Chomsky, N. (1995). *The minimalist program*. MIT.

Chomsky, N. (2000). *New horizons in the study of language and mind*. Cambridge University Press.

Chomsky, N. (2004). Beyond explanatory adequacy. In A. Belletti (Ed.), *Structures and beyond* (pp. 104–131). Oxford University Press.

Chomsky, N. (2008). On phases. In R. Freidin, C. P. Otero, & M. L. Zubizaretta (Eds.), *Foundational issues in linguistic theory: Essays in honor of Jean-Roger Vergnaud* (pp. 133–166). MIT.

Chomsky, N. (2011). Language and other cognitive systems. What is special about language? *Language, Learning and Development, 7*(4), 263–278.

Chomsky, N. (2013). Problems of projection. *Lingua, 130*, 33–49.

Chomsky, N. (2014). Minimal recursion: Exploring the prospects. In T. Roeper & M. Speas (Eds.), *Studies in theoretical psycholinguistics 43. Recursion: Complexity in cognition* (pp. 1–15). Springer.

Chomsky, N. (2019). Some puzzling foundational issues: The Reading Program. *Catalan Journal of Linguistics Special Issue,* 263–285.

Chomsky, N. (2020). Minimalism: where we are now and where we are going. Talk given at the Linguistic Society of Japan. 22 November.

Chomsky, N. (2021a). Linguistics then and now: Some personal reflections. *Annual Review of Linguistics, 7*, 1–11.

Chomsky, N. (2021b). Simplicity and the form of grammars. *Journal of Language Modelling, 9*(1), 5–15.

Chomsky, N. (2021c). Minimalism: Where we are now, and where we can hope to go. *Gengo Kenkyu (journal of the Linguistic Society of Japan), 160*, 1–41.

Chomsky, N. (2022). Genuine explanation and the strong minimalist thesis. Lecture at the Biolinguistics Fall Semester class, University of Arizona, 24 August.

Chomsky, N. (2023). Genuine explanation and the strong minimalist thesis. *Cognitive Semantics, 8*, 347–365.

Chomsky, N., Gallego, Á. J., & Ott, D. (2019). Generative grammar and the faculty of language: insights, questions, and challenges. *Catalan Journal of Linguistics Special Issue, 1*, 226–261.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. Harper and Row.

Citko, B., & Gračanin-Yuksek, M. (2021). *Merge: Binarity in (multidominant) syntax*. MIT.

Clifton, C., Fanselow, G., & Frazier, L. (2006). Amnestying superiority violations: Processing multiple questions. *Linguistic Inquiry, 37*(1), 51–68.

Collins, C. (2017). Merge(X, Y)={X, Y}. In L. Bauke & A. Blümel (Eds.), *Labels and roots* (pp. 47–68). De Gruyter.

Collins, J. (2020). Conjoining meanings without losing our heads. *Mind & Language, 35*, 224–236.

Colombo, M., & Wright, C. (2021). First principles in the life sciences: The free-energy principle, organicism, and mechanism. *Synthese, 198*, 3463–3488.

Coopersmith, J. (2017). *The lazy universe: An introduction to the principle of least action*. Oxford University Press.

Crain, S., Koring, L., & Thornton, R. (2017). Language acquisition from a biolinguistic perspective. *Neuroscience and Biobehavioral Reviews, 81B*, 120–149.

Culbertson, J., Smolensky, P., & Legendre, G. (2012). Learning biases predict a word order universal. *Cognition, 122*(3), 306–329.

Da Costa, L., Parr, T., Sengupta, B., & Friston, K. (2021). Neural dynamics under active inference: Plausibility and efficiency of information processing. *Entropy, 23*, 454.

Dalrymple, M., Kaplan, R. M., & King, T. H. (2015). Economy of expression as a principle of syntax. *Journal of Language Modelling, 3*(2), 377–412.

Dasgupta, I., & Gershman, S. J. (2021). Memory as a computational resource. *Trends in Cognitive Sciences, 25*(3), 240–251.

Dayal, V. (2017). Multiple wh-questions. In M. Everaert & H. C. Riemsdijk (Eds.), *The Wiley Blackwell companion to syntax (second edition).* Blackwell.

Dehaene, S., Al Roumi, F., Lakretz, Y., Planton, S., & Sablé-Meyer, M. (2022). Symbols and mental programs: A hypothesis about human singularity. *Trends in Cognitive Sciences, 26*(9), 751–766.

Dobashi, N. (2010). Computational efficiency in the syntax–phonology interface. *The Linguistic Review, 27*(3), 241–260.

Ehret, K., Berdicevskis, A., Bentz, C., & Blumenthal-Dramé, A. (2023). Measuring language complexity: Challenges and opportunities. *Linguistics Vanguard, 9*(s1), 1–8.

Embick, D., & Poeppel, D. (2015). Towards a computational(ist) neurobiology of language: Correlational, integrated, and explanatory neurolinguistics. *Language, Cognition and Neuroscience, 30*(4), 357–366.

Epstein, S. D., Groat, E., Kawashima, R., & Kitahara, H. (1998). *A derivational approach to syntactic relations*. Oxford University Press.

Epstein, S. D., Kitahara, H., & Seely, T. D. (2021). *A minimalist theory of simplest merge*. Routledge.

Evans, D. J., & Searles, D. J. (1994). Equilibrium microstates which generate second law violating steady states. *Physical Review E, 50*, 1645–1648.

Faul, S. (2021). Kolmogorov complexity, MATLAB Central File Exchange. Retrieved November 23, 2021.

Feldman, J. (2003). The simplicity principle in human concept learning. *Current Directions in Psychological Science, 12*(6), 227–232.

Feynman, R. P. (1972). *Statistical mechanics*. Benjamin.

Fong, S., Berwick, R. C., & Ginsburg, J. (2019). The combinatorics of Merge and workspace right-sizing. Paper presented at Evolinguistics Workshop, 25–26 May.

Fong, S., & Ginsburg, J. (2018). On constraining Free Merge. In: *The 43rd meeting of the Kansai Linguistics Society*< Konan University, Kobe, Japan.

Fox, D. (2000). *Economy and semantic interpretation*. MIT.

Frampton, J., & Gutmann, S. (1999). Cyclic computation, a computationally efficient minimalist syntax. *Syntax, 2*(1), 1–27.

Friederici, A. D., Chomsky, N., Berwick, R. C., Moro, A., & Bolhuis, J. J. (2017). Language, mind and brain. *Nature Human Behaviour, 1*, 713–722.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11*, 127–138.

Friston, K. J. (2019). Complexity and computation in the brain: The knowns and the known unknowns. In W. Singer, T. J. Sejnowski, & P. Rakic (Eds.), *The Neocortex* (pp. 269–291). MIT.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017a). Active inference: A process theory. *Neural Computation, 29*(1), 1–49.

Friston, K. J., Parr, T., & de Vries, B. (2017b). The graphical brain: Belief propagation and active inference. *Network Neuroscience, 1*(4), 381–414.

Friston, K. J., Parr, T., Yufik, Y., Sajid, N., Price, C. J., & Holmes, E. (2020). Generative models, linguistic communication and active inference. *Neuroscience and Biobehavioral Reviews, 118*, 42–64.

Friston, K. J., Sajid, N., Quiroga-Martinez, D. R., Parr, T., Price, C. J., & Holmes, E. (2021). Active listening. *Hearing Research, 399*, 107998.

Freidin, R., & Lasnik, H. (2011). Some roots of minimalism. In C. Boeckx (Ed.), *The Oxford handbook of linguistic minimalism* (pp. 1–26). Oxford University Press.

Fukui, N. (1996). On the nature of economy in language. *Cognitive Studies, 3*, 51–71.

Fuster, J. M. (2004). Upper processing stages of the perception-action cycle. *Trends in Cognitive Sciences, 8*(4), 143–145.

Gallego, Á. J., & Chomsky, N. (2020). The faculty of language: A biological object, a window into the mind, and a bridge across disciplines. *Revista De La Sociedad Española De Lingüística, 50*(1), 7–34.

Gallego, Á. J., & Martin, R. (Eds.). (2018). *Language, syntax and the natural sciences*. Cambridge University Press.

Gallego, Á. J., & Orús, R. (2022). Language design as information renormalization. *SN Computer Science, 3*(140), 1–27.

Gärtner, H.-M. (2021). Copies from "standard set theory"? A note on the foundation of minimalist syntax in reaction to Chomsky, Gallego and Ott (2019). *Journal of Logic, Language and Information*. https://doi.org/10.1007/s10849-021-09342-x

Gärtner, H.-M., & Sauerland, U. (Eds.). (2007). *Interfaces + Recursion = Language? Chomsky's minimalism and the view from syntax-semantics*. De Gruyter Mouton.

Genewein, T., & Braun, D. A. (2014). Occam's razor in sensorimotor learning. *Proceedings of the Royal Society B, 281*(1783), 2013–2952.

Gell-Mann, M., & Lloyd, S. (1996). Information measures, effective complexity, and total information. *Complexity, 2*(1), 44–52.

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences, 23*(5), 389–407.

González Escribano, J. L. (2005). Semantocentric minimalist grammar. *Atlantis, 27*(2), 57–74.

Goodman, N. (1951). *The structure of appearance*. Harvard University Press.

Goodman, N., Tenenbaum, J., Feldman, J., & Griffiths, T. (2008). A rational analysis of rule-based concept learning. *Cognitive Science, 32*(1), 108–154.

Grimaldi, M. (2012). Toward a neural theory of language: Old issues and new perspectives. *Journal of Neurolinguistics, 25*(5), 304–327.

Grimshaw, J. (2001). Economy of structure in OT. Rutgers Optimality Archive 444.

Grohmann, K. (Ed.). (2007). *InterPhases: Phase-theoretic investigations of linguistic interfaces*. Oxford University Press.

Grünwald, P. (1996). A minimum description length approach to grammar inference. In S. Wermter, E. Riloff, & G. Scheler (Eds.), *Connectionist, statistical and symbolic approaches to learning for natural language processing*. IJCAI 1995. Lecture notes in computer science (lecture notes in artificial intelligence), vol 1040 (pp. 203–216). Springer.

Grünwald, P. D. (2007). *The minimum description length principle*. MIT.

Haegeman, L. (1994). *Introduction to government and binding theory* (2nd ed.). Blackwell.

Hale, J. T. (2016). Information-theoretical complexity metrics. *Language and Linguistics Compass, 10*, 397–412.

Hauser, M., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science, 298*(5598), 1569–1579.

Hawkins, J. (2004). *Efficiency and complexity in grammars*. Oxford University Press.

Hinton, G. E., & Zemel, R. S. (1994). Autoencoders, minimum description length and Helmholtz free energy. In J. D. Cowan, G. Tesauro, & J. Alspector (Eds.), *Advances in neural information processing systems* (pp. 3–10). Morgan Kaufmann.

Hinzen, W. (2006). *Mind design and minimal syntax*. Oxford University Press.

Hohwy, J. (2016). The self-evidencing brain. *Noûs, 50*, 259–285.

Hohwy, J. (2017). How to entrain your evil demon. In T. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing 2* (pp. 1–15). MIND Group.

Hohwy, J. (2020). New directions in predictive processing. *Mind & Language, 35*(2), 209–223.

Holmberg, A., & Roberts, I. (2014). Parameters and the three factors of language design. In C. Picallo (Ed.), *Linguistic variation in the minimalist framework* (pp. 61–81). Oxford University Press.

Hornstein, N., & Pietroski, P. (2009). Basic operations: Minimal syntax-semantics. *Catalan Journal of Linguistics, 8*, 113–139.

Hutter, M. (2006). *Universal artificial intellegence: Sequential decisions based on algorithmic probability*. Springer.

Huybregts, M. A. C. (2019). Infinite generation of language unreachable from a stepwise approach. *Frontiers in Psychology, 10*, 425.

Jackendoff, R., & Wittenberg, E. (2014). What you can say without syntax: A view from evolution. In F. J. Newmeyer & L. B. Preston (Eds.), *Measuring grammatical complexity* (pp. 65–82). Oxford University Press.

Jarzynski, C. (1997). Nonequilibrium equality for free energy differences. *Physical Review Letters, 78*(14), 2690–2693.

Joshi, A. K., & Schabes, Y. (1997). Tree-adjoining grammars. Beyond WordsIn G. Rozenberg & A. Salomaa (Eds.), *Handbook of formal languages* (Vol. 3, pp. 69–123). Springer.

Juola, P. (2008). Assessing linguistic complexity. In M. Miestamo, K. Sinnemaki, & F. Karlsson (Eds.), *Language complexity: Typology, contact, change* (pp. 89–107). Benjamins.

Kaspar, F., & Schuster, H. G. (1987). Easily calculable measure for the complexity of spatiotemporal patterns. *Physical Review A, 36*(2), 842–848.

Katzir, R., O'Donnell, T. J., & Rasin, E. (2021). Introduction to the special issue on simplicity in grammar learning. *Journal of Language Modeling, 9*(1), 1–4.

Ke, H. (2019). *The syntax, semantics and processing of agreement and binding grammatical illusions*. PhD dissertation. University of Michigan.

Kirchhoff, M., Parr, T., Palacios, E., Friston, K., & Kiverstein, J. (2018). The Markov blankets of life: Autonomy, active inference and the free energy principle. *Journal of the Royal Society Interface, 15*(138), 20170792.

Kirchhoff, M. D., & Robertson, I. (2018). Enactivism and predictive processing: A non-representational view. *Philosophical Explorations, 21*(2), 264–281.

Kleene, S. C. (1952). *Introduction to metamathematics*. Van Nostrand.

Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission, 1*, 1–7.

Komachi, M., Kitahara, H., Uchibori, A., & Takita, K. (2019). Generative procedure revisited. *Reports of the Keio Institute of Cultural and Linguistic Studies, 50*, 269–283.

Korbak, T. (2021). Computational enactivism under the free energy principle. *Synthese, 198*, 2743–2763.

Korf, R. (1985). Depth-first iterative-deepening: An optimal admissable tree search. *Artificial Intelligence, 27*(1), 97–109.

Kosta, P., Schürcks, L., Franks, S., & Radev-Bork, T. (Eds.). (2014). *Minimalism and beyond: Radicalizing the interfaces*. John Benjamins.

Kush, D., Lidz, J., & Phillips, C. (2015). Relation-sensitive retrieval: Evidence from bound variable pronouns. *Journal of Memory and Language, 82*, 18–40.

Lambert, D., Rawski, J., & Heinz, J. (2021). Typology emerges from simplicity in representations and learning. *Journal of Language Modelling, 9*(1), 151–194.

Larson, B. (2015). Minimal search as a restriction on merge. *Lingua, 156*, 57–69.

Lasnik, H., & Lohndal, T. (2013). Brief overview of the history of generative syntax. In M. den Dikken (Ed.), *The Cambridge handbook of generative syntax* (pp. 26–60). Cambridge University Press.

Leivada, E., & Murphy, E. (2021). Mind the (terminological) gap: 10 misused, ambiguous, or polysemous terms in linguistics. *Ampersand, 8*, 10073.

Leivada, E., Murphy, E., & Marcus, G. (2023). DALL·E 2 fails to reliably capture common syntactic processes. *Social Sciences & Humanities Open, 8*(1), 100648.

Lempel, A., & Ziv, J. (1976). On the complexity of finite sequences. *IEEE Transactions on Information Theory, 22*(1), 75–81.

Li, M., & Vitányi, P. (2019). *An introduction to Kolmogorov complexity and its applications* (4th ed.). Springer.

Linsker, R. (1990). Perceptual neural organization: Some approaches based on network models and information theory. *Annual Review of Neuroscience, 13*, 257–281.

Lobina, D. J. (2017). *Recursion: A computational investigation into the representation and processing of language*. Oxford University Press.

Lohndal, T., & Uriagereka, J. (2016). Third-factor explanations and universal grammar. In I. Roberts (Ed.), *The Oxford handbook of universal grammar* (pp. 114–128). Oxford University Press.

Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences, 20*(10), 748–759.

Longobardi, G. (2008). Reference to individuals, person, and the variety of mapping parameters. In A. Klinger & H. Mueller (Eds.), *Essays on nominal determination: From morphology to discourse management* (pp. 189–211). John Benjamins.

Longobardi, G. (2017). Principles, parameters, and schemata: A radically underspecified UG. *Linguistic Analysis, 41*(3–4), 517–558.

Lupyan, G., & Clark, A. (2015). Words and the world: Predictive coding and the language-perception-cognition interface. *Current Directions in Psychological Science, 24*(4), 279–284.

MacGregor, J. N. (1987). Short-term-memory capacity: Limitation or optimization? *Psychological Review, 94*(1), 107–108.

MacKay, D. J. (1995). Free-energy minimisation algorithm for decoding and cryptoanalysis. *Electronics Letters, 31*, 445–447.

MacKay, D. J. (2003). *Information theory, inference, and learning algorithms*. Cambridge University Press.

Marcolli, M., Chomsky, N., & Berwick, R.C. (2023). Mathematical structure of syntactic merge. arXiv:2305.18278

Martin, A., Holtz, A., Abels, K., Adger, D., & Culbertson, J. (2020). Experimental evidence for the influence of structure and meaning on linear order in the noun phrase. *Glossa: A Journal of General Linguistics, 5*(1), 97.

McCarty, M. J., Murphy, E., Scherschligt, X., Woolnough, O., Morse, C. W., Snyder, K., Mahon, B. Z., & Tandon, N. (2023). Intraoperative cortical localization of music and language reveals signatures of structural complexity in posterior temporal cortex. *Science, 26*(7), 107223

Miestamo, M., Sinnemäki, K., & Karlsson, F. (Eds.). (2008). *Language complexity: Typology, contact, change*. Benjamins.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences, 11*(4), 143–152.

Ming, L., & Vitányi, P. (2008). *An Introduction to kolmogorov complexity and its applications* (3rd ed.). Springer.

Mitchell, M. (2009). *Complexity: A guided tour*. Oxford University Press.

Mizuguchi, M. (2019). Optional raising in ECM and labeling of XP-YP. *Generative Grammar Research, 29*(2), 373–411.

Murphy, E. (2012). *Biolinguistics and philosophy: Insights and obstacles*. Lulu.

Murphy, E. (2015a). Labels, cognomes, and cyclic computation: An ethological perspective. *Frontiers in Psychology, 6*, 715.

Murphy, E. (2015b). Reference, phases and individuation: Topics at the labeling-interpretive interface. *Opticon, 17*(5), 1–13.

Murphy, E. (2015c). The brain dynamics of linguistic computation. *Frontiers in Psychology, 6*, 1515.

Murphy, E. (2020a). Language design and communicative competence: the minimalist perspective. *Glossa: A Journal of General Linguistics, 5*(1), 2.

Murphy, E. (2020b). *The oscillatory nature of language*. Cambridge University Press.

Murphy, E. (2023a). The citadel itself: defending semantic internalism. *Global Philosophy*. https://doi.org/10.1007/s10516-023-09669-z

Murphy, E. (2023b). A future without a past: Philosophical consequences of Merge. *Biolinguistics, 17*, e13067.

Murphy, E. (2024). ROSE: A neurocomputational architecture for syntax. *Journal of Neurolinguistics, 70*, 101180.

Murphy, E., & Benítez-Burraco, A. (2018). Paleo-oscillomics: Inferring aspects of Neanderthal language abilities from gene regulation of neural oscillations. *Journal of Anthropological Sciences, 96*, 111–124.

Murphy, E., Forseth, K. J., Donos, C., Snyder, K. M., Rollo, P. S., & Tandon, N. (2023). The spatiotemporal dynamics of semantic integration in the human brain. *Nature Communications, 14*, 6336.

Murphy, E., & Shim, J.-Y. (2020). Copy invisibility and (non-)categorial labeling. *Linguistic Research, 37*(2), 187–215.

Murphy, E., Woolnough, O., Rollo, P. S., Roccaforte, Z. J., Segaert, K., Hagoort, P., & Tandon, N. (2022). Minimal phrase composition revealed by intracranial recordings. *Journal of Neuroscience, 42*(15), 3216–3227.

Narita, H. (2014). *Endocentric structuring of projection-free syntax*. John Benjamins.

Newmeyer, F. J. (2007). More complicated and hence, rarer: a look at grammatical complexity and crosslinguistic rarity. In S. Karimi, V. Samiian & W.K. Wilkins (Eds.), *Clausal and phrasal architecture: Syntactic derivation and interpretation (Festschrift for Joseph E. Emonds)* (pp. 221–242). John Benjamins.

Newmeyer, F. J., & Preston, L. B. (Eds.). (2014). *Measuring grammatical complexity*. Oxford University Press.

Palacios, E. R., Razi, A., Parr, T., Kirchhoff, M., & Friston, K. (2020). On Markov blankets and hierarchical self-organization. *Journal of Theoretical Neurobiology, 486*, 110089.

Parr, T., Da Costa, L., & Friston, K. (2020). Markov blankets, information geometry and stochastic thermodynamics. *Philosophical Transactions of the Royal Society A, 378*(2164), 20190159.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann.

Pearl, L. (2022). Poverty of the stimulus without tears. *Language Learning and Development, 18*(4), 415–454.

Penny, W. D., Stephan, K. E., Mechelli, A., & Friston, K. J. (2004). Comparing dynamic causal models. *NeuroImage, 22*(3), 1157–1172.

Piantadosi, S. T. (2021). The computational origin of representation. *Minds and Machines, 31*, 1–58.

Pietroski, P. (2005). *Events and semantic architecture*. Oxford University Press.

Pietroski, P. (2018). *Conjoining meanings: Semantics without truth values*. Oxford University Press.

Planton, S., van Kerkoerle, T., Abbih, L., Maheu, M., Meyniel, F., Sigman, M., Wang, L., Figueira, S., Romano, S., & Dehaene, S. (2021). A theory of memory for binary sequences: Evidence for a mental compression algorithm in humans. *PLoS Computational Biology, 17*(1), e1008598.

van de Pol, I., Lodder, P., van Maanen, L., Steinert-Threlkeld, S., & Szymanik, J. (2021). Quantifiers satisfying semantic universals are simpler. *PsyArXiv*. https://doi.org/10.31234/osf.io/xuhyr

Port, A., Karidi, T., & Marcolli, M. (2022). Topological analysis of syntactic structures. *Mathematics in Computer Science, 16*, 2.

Pulvermüller, F. (2014). The syntax of action. *Trends in Cognitive Sciences, 18*(5), 219–220.

Quine, W. V. O. (1995). Whitehead and the rise of modern logic. *Selected Logic Papers* (pp. 1–36). Harvard University Press.

Radford, A. (2016). *Analysing English sentences*. Cambridge University Press.

Rahman, M. S., & Kaykobad, M. (2005). On Hamiltonian cycles and Hamiltonian paths. *Information Processing Letters, 94*, 37–41.

Ramstead, M. J. D., Badcock, P. B., & Friston, K. J. (2018). Answering Schrödinger's question: A free-energy formulation. *Physics of Life Reviews, 24*, 1–16.

Ramstead, M. J., Kirchhoff, M. D., Constant, A., & Friston, K. J. (2021). Multiscale integration: Beyond internalism and externalism. *Synthese, 198*, 41–70.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2*(1), 79–87.

Rasin, E., Berger, I., Lan, N., Shefi, I., & Katzir, R. (2021). Approaching explanatory adequacy in phonology using Minimum Description Length. *Journal of Language Modelling, 9*(1), 17–66.

Reuland, E. (2011). *Anaphora and language design*. MIT.

Richards, M. (2011). Deriving the edge: What's in a phase? *Syntax, 14*, 74–95.

Rizzi, L. (1990). *Relativized minimality*. MIT.

Rizzi, L. (2001). Relativized minimality effects. In M. Baltin & C. Collins (Eds.), *The handbook of contemporary syntactic theory* (pp. 89–110). Blackwell.

Roberts, I. (2019). *Parameter hierarchies and universal grammar*. Oxford University Press.

Romano, S., Sigman, M., & Figueira, S. (2013). LT^2C^2: A language of thought with Turing-computable Kolmogorov complexity. *Papers in Physics, 5*, 050001.

Samo, G. (2021). N-merge systems in adult and child grammars: A quantitative study on external arguments. *Qulso, 7*, 103–130.

Schein, B. (1993). *Plurals and Events*. MIT.

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development, 2*(3), 230–247.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks, 61*, 85–117.

Schütze, C. T. R. (1997). *INFL in child and adult language: Agreement, case and licensing*. PhD dissertation, MIT.

Sengupta, B., & Stemmler, M. N. (2014). Power consumption during neuronal computation. *Proceedings of the IEEE, 102*(5), 1–13.

Shieber, S. (1985). Evidence against the context-freeness of natural language. *Linguistics and Philosophy, 8*(3), 333–343.

Smith, R., Parr, T., & Friston, K. J. (2019). Simulating emotions: An active inference model of emotional state inference and emotion concept learning. *Frontiers in Psychology, 10*, 2844.

Solomonoff, R. J. (1964). A formal theory of inductive inference Part i. *Information and Control, 7*(1), 1–22.

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods, 43*, 155–167.

Sprouse, J., & Almeida, D. (2017). Setting the empirical record straight: Acceptability judgments appear to be reliable, robust, and replicable. *Behavioral and Brain Sciences, 40*, e311.

Starke, M. (2001). *Move dissolves into merge*. PhD thesis. University of Geneva.

Starke, M. (2004). On the inexistence of specifiers and the nature of heads. In A. Belletti (Ed.), *The cartography of syntactic structures. Structures and beyond* (Vol. 3, pp. 252–268). Oxford University Press.

Steedman, M. (2000). *The syntactic process*. MIT.

Sternefeld, W. (1997). Comparing reference sets. In C. Wilder, H-M. Gärtner, & M. Bierwisch (Eds.), *The role of economy principles in linguistic theory* (pp. 81–114). Akademie.

Sundaresan, S. (2020). Distinct featural classes of anaphor in an enriched person system. In K. Hartmann, J. Mursell, & P. W. Smith (Eds.), *Agree to agree: Agreement in the minimalist programme* (pp. 425–461). Open Generative Syntax series. Language Science Press.

Szmrecsányi, B. (2004). On operationalizing syntactic complexity. In G. Purnelle, et al. (Eds.), *Le poids des mots: Proceedings of the 7th international conference on textual data statistical Analysis*, vol. 2 (pp. 1032–1039). Louvain-la-Neuve: Presses Universitaires de Louvain.

Tervo, D. G. R., Tenenbaum, J. B., & Gershman, S. J. (2016). Toward the neural implementation of structure learning. *Current Opinion in Neurobiology, 37*, 99–105.

Terzian, G., & Corbalán, M. I. (2021). Simplicity of what? A case study from generative linguistics. *Synthese, 198*(10), 9427–9452.

Titov, E. (2020). Optionality of movement. *Syntax, 23*(4), 347–374.

Torr, J., Stanojević, M., Steedman, M., & Cohen, S. B. (2019). Wide-coverage neural A* parsing for minimalist grammars. In: *Proceedings of the 57th annual meeting of the Association for Computational Linguistics*, Florence, Italy, pp. 2486–2505.

Tribus, M. (1961). *Thermodynamics and thermostatics: An introduction to energy, information and states of matter, with engineering applications*. D. Van Nostrand Company Inc.

Trotzke, A., & Zwart, J.-W. (2014). The complexity of narrow syntax: Minimalism, representational economy, and simplest Merge. In F. J. Newmeyer & L. B. Preston (Eds.), *Measuring grammatical complexity* (pp. 128–147). Oxford University Press.

Trudgill, P. (2011). *Sociolinguistic typology: Social determinants of linguistic complexity*. Oxford University Press.

Turing, A. (1950). Computing machinery and intelligence. *Mind, 49*, 433–460.

Ungerleider, L. G., & Haxby, J. V. (1994). ‘What’ and ‘where’ in the human brain. *Current Opinion in Neurobiology, 4*(2), 157–165.

Uriagereka, J. (2012). *Spell-out and the minimalist program*. Oxford University Press.

Urwin, S. G., Griffiths, B., & Allen, J. (2017). Quantification of differences between nailfold capillaroscopy images with a scleroderma pattern and normal pattern using measures of geometric and algorithmic complexity. *Physiological Measurement, 38*(2), N32–N41.

Vaas, R. (2001). It binds, therefore I am! Review of Rodolfo Llinás’s *I of the Vortex*. *Journal of Consciousness Studies, 8*(4), 85–88.

van Gelderen, E. (2011). *The linguistic cycle: Language change and the language faculty*. Oxford University Press.

van Gelderen, E. (2021). *Third factors in language variation and change*. Cambridge University Press.

van Rooj, I., & Baggio, G. (2021). Theory before the test: How to build high-versimilitude explanatory theories in psychological science. *Perspectives on Psychological Science*. https://doi.org/10.1177/1745691620970604

Vasil, J., Badcock, P. B., Constant, A., Friston, K., & Ramstead, M. J. D. (2020). A world unto itself: Human communication as active inference. *Frontiers in Psychology, 11*, 417.

Walkden, G., & Breitbarth, A. (2019). Complexity as L2-difficulty: Implications for syntactic change. *Theoretical Linguistics, 45*(3–4), 183–209.

Wallace, C. S., & Dowe, D. L. (1999). Minimum message length and Kolmogorov complexity. *The Computer Journal, 42*(4), 270–283.

Wexler, K. (2003). Lenneberg’s dream: Learning normal language development and specific language impairment. In J. Schaffer & Y. Levy (Eds.), *Language competence across populations: Towards a definition of specific language impairment* (pp. 11–61). Lawrence Erlbaum.

Wilder, C., Gärtner, H.-M., & Bierwisch, M. (Eds.). (1997). *The role of economy principles in linguistic theory*. Akademie Verlag.

Winn, J., & Bishop, C. M. (2005). Variational message passing. *Journal of Machine Learning Research, 6*, 661–694.

Wipf, D. P., & Rao, B. D. (2007). An empirical Bayesian strategy for solving the simultaneous sparse approximation problem. *IEEE Transactions on Signal Processing, 55*(7), 3704–3716.

Woolnough, O., Donos, C., Murphy, E., Rollo, P. S., Roccaforte, Z. J., Dehaene, S., & Tandon, N. (2023). Spatiotemporally distributed frontotemporal networks for sentence reading. *Proceedings of the National Academy of Sciences, 120*(7), e2300252120.

Yang, C., Crain, S., Berwick, R. C., Chomsky, N., & Bolhuis, J. J. (2017). The growth of language: Universal grammar, experience, and principles of computation. *Neuroscience and Biobehavioral Reviews, 81B*, 103–119.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.