

Optimality and Plausibility in Language Design

Michael R. Levot

The Minimalist Program in generative syntax has been the subject of much rancour, a good proportion of it stoked by Noam Chomsky's suggestion that language may represent "a 'perfect solution' to minimal design specifications." A particular flash point has been the application of Minimalist principles to speculations about how language evolved in the human species. This paper argues that Minimalism is well supported as a plausible approach to language evolution. It is claimed that an assumption of minimal design specifications like that employed in MP syntax satisfies three key desiderata of evolutionary and general scientific plausibility: Physical Optimism, Rational Optimism, and Darwin's Problem. In support of this claim, the methodologies employed in MP to maximise parsimony are characterised through an analysis of recent theories in Minimalist syntax, and those methodologies are defended with reference to practices and arguments from evolutionary biology and other natural sciences.

Keywords: biology; evolutionary plausibility; Minimalism; philosophy of science; syntax

1. Introduction

There is no point in using the word 'impossible' to describe something that has clearly happened.
(Douglas Adams)

The Minimalist Program (henceforth, often referred to as Minimalism or simply MP) in generative syntax has been the subject of much rancour, a good proportion of it stoked by Chomsky's suggestion that "language design may really be optimal in some respects, approach[ing] a 'perfect solution' to minimal design specifications" (Chomsky, 2000a: 93). A particular flash point has been the application of Minimalism to speculation about how language evolved in the human species, most prominently represented by the Merge-only hypothesis in generative syntax (Chomsky, 2000b) and the saltationalist claims often made in parallel (Hauser *et al.*, 2002). To date, Anna Kinsella (Parker) has carried out the most extensive investigation into how well motivated Minimalism may be in relation to the evolution of human natural language syntax (Parker 2006; Kinsella, 2009,

I'd like to thank Kleantes K. Grohmann for his guidance and generosity, and also three anonymous reviewers for their very helpful criticisms and comments.



2015; Kinsella & Marcus, 2009), undertaking to look at “what we know from evolutionary biology about what typically evolving systems look like, what kinds of properties they have, and then applying this to questions about the plausible nature of language” (Kinsella & Marcus, 2009: 187). The conclusion is a strongly dissenting one, claiming that a more suitable approach “may reverse this [Minimalist] trend, and look towards possible imperfections as a source of insight into the evolution and structure of natural language” (Kinsella & Marcus, 2009: 207). The vote of *evolutionary plausibility*, it is claimed, counts against Minimalism.

This paper presents the countering view that what we know about biological design—and the kinds scientific inference needed to explain it—substantiate Minimalism as a plausible evolutionary hypothesis. Towards this end, section 2 makes some clarifications about the methodology and objectives of Minimalist syntax and introduces some technical language for discussing the virtues of Minimalism as a metric of evolutionary plausibility. In sections 3 and 4, I characterise the methodologies employed in MP through an analysis which exemplifies the use of redundancy, economy, and efficiency in Minimalist syntax. Building on this characterisation, sections 5 and 6 mount a defence of those methodologies with reference to practices and arguments drawn from contemporary evolutionary biology and neighbouring natural sciences.

2. Optimality and Evolution

2.1. ‘The Best of All Possible Language Faculties’

In the following passage, Kinsella and Marcus lay out an argument against the Minimalist conception of language evolution.

[A]t least one strand of recent linguistics—its tendency towards a presumption of perfection—is at odds with two core facts: The fact that language evolved quite recently (relative to most other aspects of biology) and the fact that even with long periods of time, biological solutions are not always maximally elegant or efficient. To our minds, anyway, the presumption of perfection in language seems unwarranted and implausible [...]. (Marcus & Kinsella, 2009: 207)

A plausible account of language evolution, they claim, leaves scant margin for optimal design. They consider the following metrics against which one could assess this claim:

Language might be considered optimal if communication between speaker and hearer were as efficient as possible. [...] Another possible measure of optimality might be in terms of the amount of code that needs to be transmitted between speaker and hearer for a given message that is to be transmitted. [...] Could language be a system that yields an optimal balance between ease of comprehension and ease of acquisition? (Kinsella & Marcus, 2009: 196)

It is clear from these speculations that the notion of perfection under consideration takes optimal *communication* to be the relevant metric. A casual examination of the range of biological traits provides *prima facie* confirmation of Kinsella & Marcus’ (2009) scepticism: The biological world is teeming with messy,

unlikely solutions to environmental pressures, an observation which undergirds Kinsella and Marcus' well-founded conviction that language *qua* communicative system is more akin to Rube Goldberg machine—or a 'Kluge' in Marcus' (2009) terms—than a precision-engineered device.

The Minimalist conception of optimal design, however, is fundamentally different insofar that the faculty of language (FL) is not a communicative system—or a 'functional' system of any kind—but rather FL is a theory of a physical object. A more appropriate comparison is Turing's well-known study of morphogenesis which explains biological design by appealing to necessary interactions of matter—what neurobiologists Reeve & Sherman (2001: 64f.) referred to as "the surprisingly ordered of simple underlying processes". Optimality in the functional sense is quite distinct to optimality in the latter, developmental sense. There is no contradiction, for instance, in the design of zebra stripes being sub-optimal with respect to its function as camouflage yet also highly optimal as a solution to the developmental (i.e. biochemical) gully that must be breached to bring about this evolutionary novelty. The question of interest to Minimalists is to "what extent language is a 'good solution'" to the conditions imposed by other cognitive systems with which language interacts (Chomsky, 2000a: 9). This latter conjecture is in keeping with the Minimalist hypothesis that much of human language design can be explained by the introduction of a hierarchical form of structure to an existing "conceptual-intentional" cognitive system (roughly, the faculty of thought) and its externalisation through a sensori-motor system (roughly, the capacity for producing sound); which is to say, syntax is *for thought* in the sense that its structure was largely determined by the constraints of a pre-existing conceptual-intentional cognitive faculty.

In the context of language evolution, then, optimality is a *causal* hypothesis about how our changing biology has structured cognitive systems with respect to one another, and not a normative claim about the adaptive value of cognitive traits. The statement "even with long periods of time, biological solutions are not always maximally elegant or efficient" thus represents a departure both from the Minimalist conception of FL as an instance of biological design and from the Minimalist conception of optimality as a causal rather than normative (adaptive or functional) metric.¹ This latter notion of optimality recalls the Leibnizian form of optimism proffered by computational neuroscientist Cherniak to describe the maximally efficient component placement that characterises the human brain: the human language faculty represents the "best of all possible language faculties" (quoted in Chomsky, 2005: 6, Cherniak's actual phrase is "the best of all possible brains" 1995: 522; see also section 6 below). Kinsella & Marcus' (2009) criticisms on the basis of the communicative efficacy of language thus rebut a misconstrued version of the Minimalist conception of optimality.

¹ Kinsella does briefly give a more accurate portrayal of Minimalist desiderata in other places. For instance, she and Marcus argue that "it is unrealistic to expect language to be a perfect or near-perfect solution to the problem of mapping sound and meaning, and equally unrealistic to expect that all of language's properties can be derived straightforwardly from virtual conceptual necessity" (Kinsella & Marcus, 2009: 203).

2.2. *Darwin's Problem and Parsimony*

The question immediately posed when adopting this understanding of optimality is: What makes one theory of Narrow Syntax more optimal than any other theory?² A simple gloss to the Minimalist conception of optimality is what philosophers of science have taken to calling 'parsimony' (Popper, 1959; Simon, 1969; Kitcher, 1976; Sober, 2015)—the kind of simplicity and elegance that is typical of good scientific theories in all of the natural sciences. One aspect of parsimony which has arisen in the context of language is what has been dubbed 'Darwin's Problem' (Boeckx, 2009: 45): Postulating a large number of events resulting in FL is almost certainly inappropriate given the short space of time available and Darwin's Problem therefore militates for a saltationalist account of language; in other words, an account in which the novel language phenotype emerged rapidly with only a few evolutionary events.

2.3. *Three 'Optimalities'*

With these clarifications in mind, it will be useful to introduce some terminology for understanding how the different claims of linguists, cognitive neuroscientists, and evolutionary biologists can fit together to form a clearer picture of what Minimalism could mean as a theory of linguistic evolution. A well-established distinction in the Minimalist literature is that between methodological and substantive minimalism. The former, Chomsky notes, has a merely "heuristic and therapeutic value" (Chomsky, 2000b) for enquiry. It is methodological insofar that its motivation is not unique to linguistics—it is a general principle of science—and in that it does not rely on any ancillary hypotheses about the structure of the world. Substantive minimalism, contrastively, is the extent to which the causal hypothesis outlined in section 2.1 above is true of language. An example of substantive minimalism which I will elaborate on below is the apparently pervasive phenomenon of 'least effort' principles in syntax. The conclusion to Darwin's Problem reached by Minimalists, quite opposite to that reached by Kinsella & Marcus (2009), is that, because there is a great deal of phenotypic change to be explained in only a short span of evolutionary time, it must be assumed that something "comes for free", or is given *a priori*, to explain the dramatic variation. There is an obvious analogy here between the form of Darwin's Problem and that of the wellspring of generative metatheory, the poverty of the stimulus argument (or 'Plato's Problem'): The structure of FL is *underdetermined* by the environment, similar to the circumstance encountered by the child learner, because of the insufficient time and environmental resources available to ensure the correct final state emerges.

² Narrow Syntax represents one half of the distinction made in Hauser *et al.* (2002) between the faculty of language *broadly* conceived, and the faculty of language *narrowly* conceived. The former denotes every aspect of FL which is *sufficient* for human language—the presence of a tongue, the ability to distinguish sounds of the appropriate length and quality, and so on. The latter is a subset of the first, denoting only the aspects of FL which are *uniquely necessary* for language. That is to say, Narrow Syntax is the computational system which differentiates human language from other linguistic traits common to non-linguistic (and therefore also non-human) forms of cognition.

The methodological and substantive motivations for Minimalism are equally important to the enterprise and converge on similar theoretical objectives. Crucially, however, the two are different in their justifications. It must be recognised that the optimality of the *physical/biological object* ‘language’ is a distinct proposition to the optimality of the formalisms making up the *theory* of the physical/biological object ‘language’ and that this in turn is a distinct proposition to the simplicity of the *causal-historical sequence of events* which resulted in the design of language. Though related, these are each distinct propositions that pertain to different kinds of scientific inference. The first of these propositions is a claim about the organisation of a physical structure in the world—the question is whether or not nature is capable of producing (structurally) optimal biological traits. We may call this doctrine *Physical Optimism*. A second prong of parsimony, which we can contrastively dub Rationalist Optimism, contends that redundancy is undesirable in theories on epistemological rather than purely empirical grounds. We may designate as *Rational Optimism* any supra-empirical principle of scientific theory selection that is not an ontological commitment about the nature of the physical world.³ The last of these propositions, constituting a resolution to Darwin’s Problem, will henceforth be referred to as Causal–Historical Optimism.

We can distinguish, then, three justifications for parsimony which may figure into the plausibility of an evolutionary account of language design: Rational Optimism, Physical Optimism, and Causal–Historical Optimism. It must be noted that the three are not entirely independent; a physically optimal language faculty (the biological object) obviously increases the plausibility of a saltationist approach to Causal–Historical Optimism because a physically optimal language faculty is easier for evolution to reach. Similarly, a parsimonious biological object will naturally lend itself to the existence of an optimal theory of language. These connections are explored further in section 5 and section 6 below.

3. The Explananda of Minimalism

3.1. Parsimony and ‘Principled Explanation’

In addition to establishing that parsimony is a virtue for explaining language design, it must also be shown that MP is in fact a parsimonious theory in the appropriate ways. Here it is important to accurately characterise the methodology and objectives of syntactic Minimalism. One of the main objectives established in the Minimalist literature is the need to provide a ‘principled’ explanation for the properties of language with the corollary that any theoretical posits which are not principled ought to be considered suspect. A property of language, according to

³ Bar for a single perfunctory (though, strangely, not wholly dismissive) mention of this epistemological aspect of scientific parsimony, Kinsella and Marcus never satisfactorily address its significance: “In one respect, this notion [of parsimony in syntax] is admirable (if unsurprising): Linguistic theorizing, like all scientific theorizing, should be guided by considerations of parsimony” (Kinsella & Marcus, 2009: 199).

Chomsky, can be considered principled insofar that it “can be reduced to [1] the third factor and to [2] conditions that language must meet to be usable at all” (Chomsky, 2005: 10; numerical annotations mine).

The ‘third factor’ is a somewhat enigmatic reference to elements of what Cherniak and others have termed non-genomic nativism—that is, aspects of biological design which follow from geometrical and computational necessities and are thus neither inherited nor acquired.⁴ The second element of principled explanation, “virtual conceptual necessity”, refers simply to the virtue of building theories from first principles and abandoning unnecessary theoretical machinery. The boldest formulation of Minimalist syntax, the so-called Strong Minimalist Thesis, is based on the hypothesis that FL minimally satisfies the requirements of (1) the third factor and (2) virtual conceptual necessity. The task of Minimalist syntax, then, is to determine which elements of the theory are minimally satisfying—that is, which are necessary—and to achieve as much empirical coverage of the relevant facts of language as possible using only these elements plus those which can reasonably be derived from the third factor.

In practice there are three basic categories of parsimony used in MP. The first implores us to make maximal use of existing explanatory technology to explain facts. The motivation here is clear enough—the reduction of explanatory *redundancy* is the salient virtue. The second strategy is to use the minimal technology necessary to explain the requisite facts, what we may call the *economy* of explanatory technology. The first two of these are two sides of the same coin which I will call *unification* for obvious enough reasons. The third maxim is to assume a general condition of computational *efficiency* in computation. Below I introduce three simple and fairly uncontroversial syntactic explananda—discrete infinity, displacement, and binding theory—and in section 4, I demonstrate how MP applies the desiderata of redundancy, economy, and efficiency to derive a more parsimonious theory of these explananda.

3.2. *Discrete Infinity*

One of the earliest discoveries pertaining to the formal properties of human natural languages was that they do not belong to the class of regular languages which can be generated by a finite-state machine (Chomsky, 1956, 1959). A finite-state machine is an abstract formal device, essentially a more restricted Turing machine, which consists of an input, a set of states, and a set of rules for changing state based on the input. Finite-state machines generate only a subset of the possible languages; more powerful abstract devices, which differ principally in their capacity to ‘remember’ strings from the input, are required to generate the full set of possible languages, including human natural languages. As an illustrative point, Berwick *et al.* (2011) have shown that the song of Bengalese finches *can* be generated by a finite-state machine and consequently belongs to the class of regu-

⁴ An example I will elaborate on in section 6 is the structure of neural arbors which are optimally spatially arranged, not because of a process of adaptive design, but because of a geometrical necessity shared by physical phenomena of numerous scales and origins—branching rivers, crystalline structures, and so on.

lar languages. Finch song conforms to this pattern as it contains sequences of notes repeated and reused throughout the duration of the song, but never reuses these sequences inside other sequences (Berwick *et al.*, 2011: 115; see Figure 1).

A finite-state grammar without dependencies can be represented as:

$$a^n b^m$$

What Chomsky showed in the mid-1950s is that, unlike finch song, human language (or, really, the English language) belongs to a larger set of languages that can contain dependencies. Unlike finite-state grammars, the dependencies contained in human languages require the ability to shift the value of a string onto a ‘stack’ while a second string is being processed and recall it at a later point. This capacity for memory is captured by the formalism of a push-down stack automaton.⁵ What this means is that a valid string can be, for instance, a sequence of *as* followed by the same number of *bs*, a string which is mirrored (*aaabbb-bbbaaa*), repeats itself (*aaabbb-aaabbb*), and so on, as represented in the following abstract grammar:

$$a^n b^n$$

This fact is evident in English when sentences are of the kind ‘If S_1 , then S_2 ’. Strings of this type cannot be generated by finite-state machines because a string in S_1 may depend on a string arbitrarily distant to it. In the string in (1), for instance, the verb in S_2 must agree in number with the subject of S_1 .

(1) If [_{S_1} the boy_a gets the girl] then [_{S_2} he is_a happy]

The resulting dependency looks like ‘ $a^1 b^2 \dots b^2 a^1$,’ a subset of those generated by a *context-free* grammar.

This basic characteristic has returned to prominence in recent discourse framed as discrete infinity. Discrete infinity, the Minimalist claim goes, marks a *sui generis* property of human cognition insofar that the capacity to generate hierarchically arranged combinations of discrete units constitutes a larger subset of the set of possible languages than any organisation of the discrete units alone could produce. Human language is thus formally distinct to the communication systems of other species.

⁵ Grammars which can remember more than one value—taking the form $a^n b^m c^n$ —are *context-sensitive*. There is some evidence that human languages are mildly context-sensitive, for instance in ‘such that’-sentences (Higginbotham, 1984):

(i) The girl_a such that the dog ran from her_b to him_c sat down on the bench.

Whether context-sensitivity is a substantive aspect of linguistic cognition or merely an artefact of domain general processing is unclear.

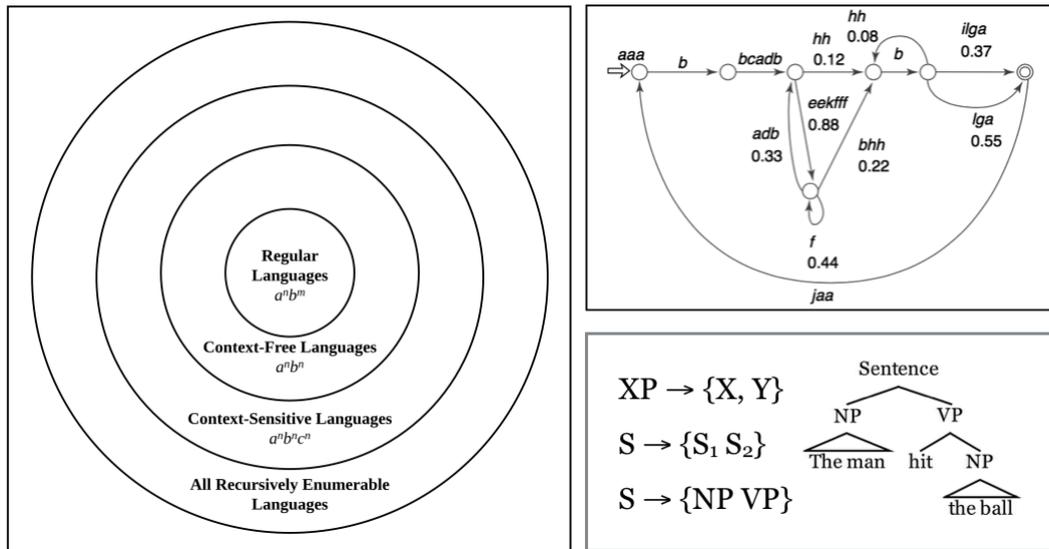


Figure 1: Formal properties of human and non-human language. Left: The Chomsky hierarchy; each category of languages is a subset of those generable by the larger set. Top right: A finite-state grammar of the Bengalese finch. Letters indicate song notes and numbers indicate probabilistic state transitions (Berwick et al., 2011: 117). Bottom right: phrase structure rules and a sentence in a dependency grammar.

3.3. Displacement

A second phenomenon unique to human languages is that lexical items are often interpreted semantically in a position different from that of their phonological expression. This *displacement* effect is readily discernible in what have traditionally been considered to be the product of transformations in a covert (i.e. phonologically unpronounced) level of syntax—D(eep)-structure in Chomsky (1981). As an example of displacement, consider the sentences in (2) and (3):

(2) Children hate broccoli.

Semantically, this sentence states:

(2') Gen x (child (x): hates broccoli (x))

Or “Typically, for x s such that x is a child, x hates broccoli”. When a question is formed from this proposition we get:

(3) What do children hate t ?

The semantic proposition expressed by the sentence is “For what x is it the case that children hate x ?”. In this case, the unknown element x does not appear adjacent to the verb *hate*, as *broccoli* does in (1), but rather it appears adjacent to *do* in the form of the pronoun *what*. Our semantic interpretation is nonetheless that children hate x , as indicated by the paraphrase “Children hate what?” Displacement, then, is the idea that elements like *what* are interpreted twice—in this case, once as a subject of the verb *do* and again as the object of the verb *hate*.

3.4. Binding Theory

The final explanandum to be treated here is binding theory, which aims to explain the distribution of co-indexed nominals. The basic data are shown below:

- (4) [Mary's father]_i hated himself_i.
- (4') * [Mary_i's father] hated herself_i.

- (5) John_i saw him_j.
- (5') * John_i saw him_i.

- (6) Jane_i saw Jane_j.
- (6') * Jane_i saw Jane_i.

The sentence pair in (4) shows that the reflexive anaphor *him-/herself* may not be co-indexed with antecedent inside a genitive phrase. Similar relationships hold for (5)–(6) where the co-indexed nominals and possible interpretations of indexing are strictly limited in grammaticality. The key insights are that the distribution of co-referring nominals is closely related to (i) the *locality* of an antecedent and (ii) the antecedent being *c-commanded* by the element with which it is co-indexed.

Locality here refers to the notion of belonging to the same 'domain' where a domain may be constituted by a phrase boundary. C-command determines the relationship between the antecedent and the anaphor (see the simplified tree structures in Figure 2). In (7), for instance, *John* is both co-indexed with and c-commands the reflexive anaphor *himself*, but the unacceptability of (7) is a result of the reflexive not being 'local enough' to its antecedent.

- (7) * [_β John_i thinks [_α that Mary saw himself_i]]

That is, because β —and not α —is the binding domain of John, *himself* is not bound in its domain and ought to take a pronominal.

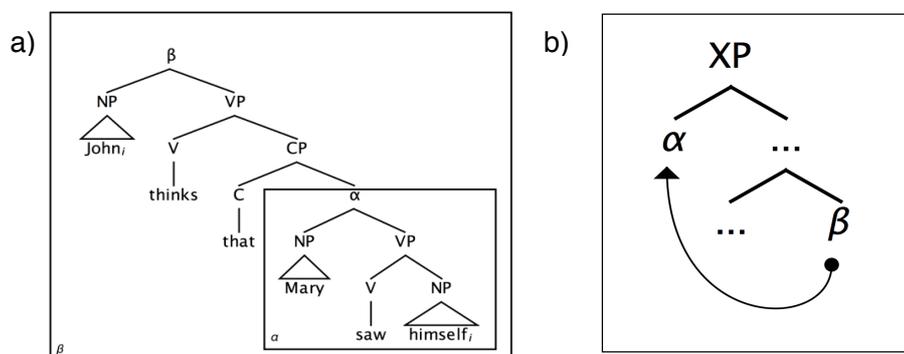


Figure 2: The binding domain and c-command. (a) John c-commands and is co-indexed with himself, but the domain of John is β , not α . (b) α c-commands β when the phrase containing α —XP in the above tree—contains β or any phrase containing β .

3.5. Summary

The picture of FL suggested by the above—though far from factually or historically complete—is one of several highly distinct formalisms explaining what appear to be quite heterogenous axiomatic systems. *Prima facie*, this heterogeneity suggests that there must be numerous historical-causal events, each responsible for the distinct formal properties of language. In the section to follow, the MP practices of redundancy, economy, and efficiency will be demonstrated with respect to four of these systems: phrase structure rules, transformations, c-command, and the notion of a binding domain.

4. Minimising Syntax

4.1. The Objectives of Minimalism

The Minimalist conjecture is that at least some of these formalisms must be eliminated if an evolutionarily plausible account of FL is to be given. This section exemplifies the methodologies of redundancy, economy, and efficiency as they are applied to reaching the goal of a plausible FL. The aim is to articulate the kind of desiderata Minimalism employs in accounting for the above linguistic facts in a maximally parsimonious way.

4.2. The Merge-Only Hypothesis

The strongest, and possibly most controversial, theory to have emerged from MP is the Merge-only hypothesis which proposes that Narrow Syntax is constituted by a single computational operation, MERGE.⁶ This conjecture is made on the grounds that MERGE is a virtually necessary component of any computational system which can generate a non-finite set of strings (i.e. a system capable of producing an unbounded array of embedded strings); *any* computational operation responsible for the dependencies ubiquitous in human languages, the claim goes, will require an operation which embeds an object within another object and this operation can be abstractly described as MERGE. Thus, the significant claim of MP is that MERGE is conceptually *necessary*, not merely conceptually *sufficient*. The methodological tenet of *redundancy* requires that all other conceptual apparatuses in the theory should be considered suspect, and the methodological tenet of *economy* requires that this virtually necessary component should be employed for maximal explanatory coverage. The Merge-only hypothesis is a clear demonstration of a unification which achieves both a reduction in redundancy and a maximal use of economy. MP is largely an exercise in making maximal use of MERGE, as well as some efficiency assumptions which are attributed—enigmatically, as it stands—to the third factor, again in line with the definition of principled explanation given in section 2.

⁶ As an anonymous reviewer points out, Merge-only is something of a misnomer if Merge alone is not capable of satisfying the legibility conditions of the sensori-motor and conceptual-intentional systems. The role of interface conditions is more central in recent Minimalist proposals.

Chomsky (2000b) presents a theory of MERGE which accounts for both the unbounded character described in section 2.1 and the displacement effect described in section 2.2. However embedding is achieved, the Minimalist claim goes, that operation must resemble the abstract computation MERGE such that:

$$\text{MERGE}(\alpha, \beta) \rightarrow K = \{\alpha, \beta\}$$

Where α and β are lexical items drawn from the lexicon and K is a new syntactic unit formed by applying MERGE to α and β . This new complex syntactic object K can then be MERGED with another syntactic object, so that:

$$\text{MERGE}(\gamma, K) \rightarrow \{\gamma \{\alpha, \beta\}\}$$

Unlike finite-state grammars, in a grammar of this kind the new objects can grow in length to become complex strings, which in turn can be MERGED with other complex strings. Returning to (1), reprinted here as (8):

- (8) If [_{s₁}the boy gets the girl] then [_{s₂}he is happy]

In terms of the abstract formal characteristics of human languages, sentences like (1)—with the dependency structure ‘ $a^1b^2 \dots b^2a^1$ ’—can be accounted for because the values of complex strings like S_1 and S_2 can be ‘remembered’ as a complex, merged whole as captured by the push-down stack formalism employed by Berwick *et al.* (2011: 120).

4.3. *The Copy Theory of Movement*

Recall that displacement involves a lexical item interpreted in two positions as shown in (2) and (3), reprinted here as (9) and (10).

- (9) Children hate broccoli.
“For x , x s hate broccoli.”
- (10) What do children hate t ?
“For what y , x s hate y .”

The mystery is that, in the second sentence, the semantic interpretation of y appears to occur in both object and sentence initial position. A natural exposition may claim that an operation is acting on y , shifting it upwards. Call this second operation MOVE. This would account for displacement, but at the cost of the additional stipulation of a second operation. The account in Chomsky (2000b) forwards an argument to the effect that MERGE can account for the same explananda as MOVE and is thus methodologically preferable. This unification is possible, he claims, if it is assumed that MERGE can apply both to new objects drawn from the lexicon, as outlined above, but also to objects already inside the merged syntactic object.

This latter version of MERGE operates in the following way:

$$\text{MERGE}(\alpha, K) \rightarrow \{\alpha \{ \gamma \{ \alpha, \beta \} \}\}$$

That is, MERGE calls the object α which is already merged within the object $\{\gamma \{ \alpha, \beta \}\}$ thus making α the new head of the object $\{\alpha \{ \gamma \{ \alpha, \beta \}\}\}$. We can distinguish EXTERNAL MERGE, where the two objects MERGED are different, from this operation of INTERNAL MERGE, where one of the objects MERGED is internal to the complex object. If one of the internally merged objects is not phonologically pronounced, this will give the appearance of α having ‘moved,’ as indicated in Figure 3. This is sometimes referred to as the copy theory of movement. As outlined in Chomsky (2007), ‘copy’ is a *façon de parler* and not a bona fide operation; the two identical instances of α results simply from MERGE applying internally in line with the most economic principle, namely that neither object is altered by the operation of MERGE (the so-called ‘No Tampering Condition’).

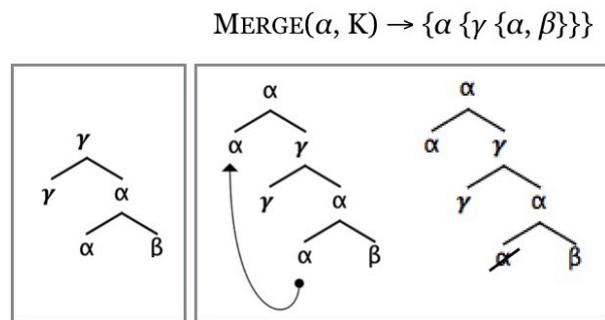


Figure 3: The copy theory of movement. Displacement and discrete infinity can be explained with a single unified theory if we assume that MERGE can apply to objects within the syntactic structure.

The copy theory of movement is an example of unifying parsimony in that two conceptual technologies (MOVE and MERGE) have been subsumed under a single, more encompassing one, thus eliminating redundancy.

4.4. C-Command and INTERNAL MERGE

An attempt to account for the technological complications of c-command and minimal domain which exemplifies the Minimalist method is that of Hornstein (2001).⁷ Recall the breakdown of binding theory in section 2 above, particularly that the distribution requires that (i) the *locality* of an antecedent and (ii) the antecedent being *c-commanded* by a co-indexed element. The two key technologies which must be explained are thus the notion of a ‘domain’ and the formal notion of c-command. Hornstein (2001) claims that the copy theory of movement makes sense of binding without these ‘messy’ stipulations:

⁷ As an anonymous reviewer points out, Hornstein’s proposals importantly rely on the existence of the operations Move and Copy—for instance, to account for Improper Movement restrictions. I’ve overlooked these inconsistencies here, as they go well beyond the scope of this paper, and continued the discussion in terms of the copy theory of movement.

[The Minimalist Program] already has a notion of local domain, i.e., ‘minimal domain,’ as part of its theory of movement. [...] Standard considerations of theoretical parsimony would favor eliminating one of these locality notions.

(Hornstein, 2001: 153)

For instance, if INTERNAL MERGE is applied to an element α of an XP, the following will result:

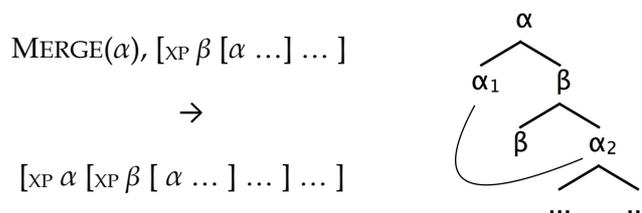


Figure 4: α_1 will c-command α_2 as a result of INTERNAL MERGE.

In the copy theory, the higher α will form a ‘chain’ (in GB parlance; see Chomsky, 1981) with the lower α simply because they are the same element copied via INTERNAL MERGE. The c-command relation emerges from the requirement that α attach to the root node purely because any element MERGING with an internal element will be dominated by a phrase dominating the internal element (see Figure 4) and the economy condition Shortest Move is capable of explaining the requirement for locality. This is because, for example, in (11) MERGING *John* to sentence initial position would violate it.

- (11) * $[\text{John}_i \text{ thinks } [\text{that Mary saw himself}_i]]$

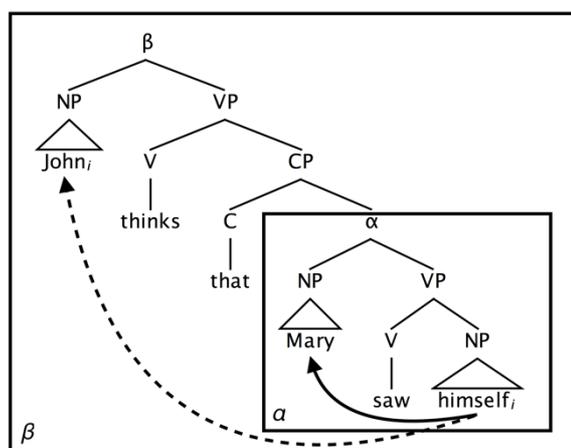


Figure 5: Shortest Move can account for the local domain of binding theory.

Hence, Hornstein’s approach to binding is an exemplification of how the requirements of binding theory can be met with a Minimalist methodology employing no greater technological complication than MERGE and Shortest Move.

4.5. Summary

Efficiency conditions are not motivated in the same way MERGE is—it is not a conceptual necessity that language is computationally efficient. Proceeding from

the virtual conceptual necessity of MERGE, MP unifies the theoretical machinery of three distinct formal systems under this single mechanism. Discrete infinity, displacement, and c-command—probably the most formally conspicuous aspects of human natural language syntax—can be accounted in a near maximally unified theory of syntax. The additional of Shortest Move further derives a local domain restraint which has applications in many areas of syntax and is exemplary of the use of efficiency. The dictum of efficiency can therefore still make a claim to being parsimonious if it can unify numerous heterogeneous conceptual technologies under a single mechanism.

5. Rational Optimism

5.1. *Two Justifications for Rational Optimism*

Rational Optimism represents an *epistemologically* motivated justification for parsimony based on the conjecture that simpler theories are *ipso facto* more likely to be true theories. Contrary to Kinsella, an evolutionary story with fewer mutations is in fitting with evolutionary biological practice. One candidate is a probabilistic or ‘frequentist’ approach to causal inference which validates the intuitive assumption that a single common cause of two events is more plausible than multiple independent causes. The frequentist interpretation of parsimony lends itself particularly well to the saltationist hypothesis for language since it is part of the central methodology of cladistics, the science of evolutionary history. A second approach differs from the first in that it takes as its focus the possibility of error in scientific models rather than the likelihood of causes. This is a strong justification of unification as a methodology and is routinely used in the natural sciences to reduce the level of error in modelling by estimating the ‘overfit’ of a theory with respect to an impoverished data set. Both of these justifications, if correct, constitute a supra-empirical principle of parsimony that support MP’s methodology and hypotheses.

5.2. *Likelihood and Parsimony*

Let’s proceed with the first, frequentist, interpretation of plausibility and evaluate its utility to the notion of parsimony as it pertains to language evolution. An example of how parsimony may increase likelihood, taken from Sober’s (1988) discussion of Reichenbach, can get us on our way.

Given a pair of correlated facts—say, both my and my neighbour’s car doors being scratched—a simple explanation for the pair may be that my neighbour has dinged my car door with his or her own car door thus causing the damage to both door simultaneously. Call this hypothesis E_1 . A more complex explanation, requiring the postulation of more agents and causes, is that a third neighbour dinged both of our car doors independently. Call this E_2 .

(E_1) Neighbour 1 damages both his/her car door and my car door at the same time.

(E_2) Neighbour 2 damages Neighbour 1’s door and my door in two separate incidents.

Assume both neighbours have an equal probability of damaging my car door—they will each do so around once a year, giving them each an approximately 0.3% probability of having damaged my car door on any particular day this year—and that instances of car door damage nearly always result in both doors being damaged—about 90% of the time. E_1 , as well as being simpler, confers a much higher likelihood on the outcome.⁸

5.2.1. *Conjunctive Forks*

Following the logic of Sober's discussion of the notion of a 'conjunctive fork,' as formulated by Reichenbach (1956), we can get some initial purchase on why likelihood improves with parsimony. We may say that a correlation between events is *probabilistically dependent* when one conspicuously co-occurs with another. A correlation occurs, then, in the case that:

$$Pr(A_1|A_2) > Pr(A_1) \times Pr(A_2)$$

Which is to say, correlation occurs when the observed probability of two events— A_1 and A_2 —occurring together is greater than the observed probability of them occurring independently. Having knowledge of A_1 therefore gives us probabilistic knowledge of A_2 which the probabilities of each event alone would not reveal. In the above case of the damaged car door, there is a very high probability of damage being caused to both doors involved. Call this probability T . It was also the case that the cars in question were neighbours, so when the probability of damage is present for A_1 it is also present for A_2 . Call this assumption C .

The interesting fact which Reichenbach noted is that if causal hypothesis E_1 is assumed then the presence or absence of T under the assumption of C is sufficient for us to estimate the probability of both doors being damaged. It is no longer necessary to posit the dependence of one event on the other because knowing the probability of damage and the fact that the two agents are neighbours explains the correlation. If we posit a joint cause of A_1 and A_2 in this way, Reichenbach claims, we have a conjunctive fork: A postulated cause which renders the probabilities of two correlated events independent. What we have described here, then, is a way of understanding why postulating common causes—that is, postulating *fewer* causes—leads to better explanations. This 'Principle of the Common Cause' claims that positing fewer causes for the same net effect will, *ceteris paribus*, deliver a better explanation.

5.2.2 *The Principle of the Common Cause and Cladistic Parsimony*

Sober provides a succinct example of the utility of parsimony to historical inference which will make the association clearer. Sober has in a mind a particular

⁸ We can see how this is so arithmetically by attending to the probabilities of E_1 and E_2 respectively. The probability of one neighbour damaging my door, plus the probability that both doors will have evidence of having being damaged is $Pr(.003)(.9)$. The probability of E_1 is therefore greater than the probability that Neighbour 1 and my car doors will have each been damaged by Neighbour 2 independently, which is $Pr(.003)(.003)$.

problem of historical inference, namely the principle of cladistic parsimony. As the name suggests, cladistic parsimony the similarities between species is best explained by positing common ancestry wherever possible. This precept has an inverse: As well as maximising the number of posited common derived characters,⁹ we should minimise the number of posited homoplasies—parallel, or convergent, similarities which have evolved independently.

This virtue can be demonstrated by taking a simple case like that in Fig. 6, where each branch—A, B, and C—represents a species.

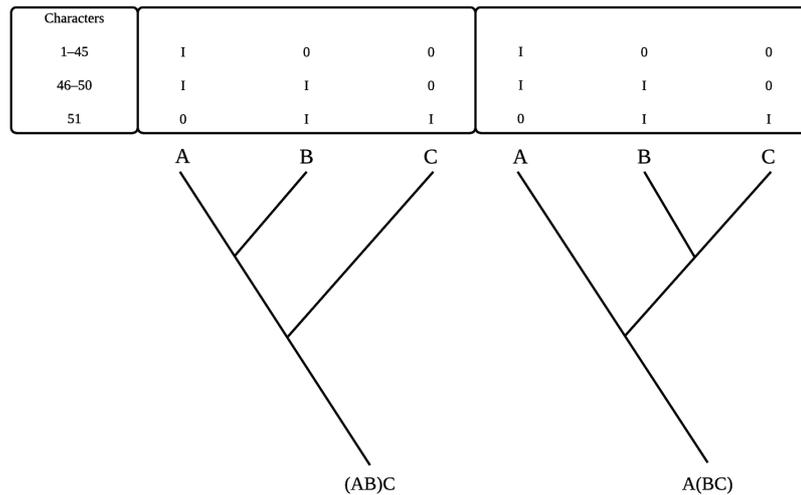


Figure 6: Cladistic parsimony. Two possible ancestries (bottom) for a set of characters (in table, above). Adapted from (Sober, 1988: 30).

Faced with the problem of reconstructing the ancestry of a character will mean deciding whether any two of A, B, and C have a common ancestor which the other lacks. The only evidence available to us is the presence or absence of various characters, as represented in the table at the top of the figure and we further assume that all three species have at least one common ancestor. The problem here is that, while positing the ancestry depicted in the figure on the left perfectly explains the distribution of characters 1–45 and 46–50, we must then assume that B and C each evolved character 51 independently. By contrast, the ancestry depicted in the figure on the right explains the homology in characters 1–45 and 51, but we must assume that A and B each independently evolved characters 46–50. According to cladistic parsimony, the *best* theory of the historical ancestry of these characters is therefore (AB)C, as it posits the fewest homoplasies.

⁹ A tangential note on the terminology of cladistic analysis: What I have simply called a ‘derived character’ (i.e., any non-zero character in the figure) is a vernacular term for an apomorphy. Any apomorphy which is inherited from a direct common ancestor (A and B in the figure on the left) is a synapomorphy. By extension, what I have called an ‘ancestral character’ is a plesiomorphy (all zero-valued characters in the figure) which, when shared, become symplesiomorphies (B and C in the figure on the right). Since my discussion of cladistic parsimony is less central than Sober’s, I stick with the less jargonistic ‘derived’ and ‘ancestral’ character, using ‘shared’ or ‘common’ to indicate homology.

Increasing the resolution from many clades and their respective characters to a single species (*Homo sapiens*) and its characters should not alter the conclusions drawn from Sober's reasoning: If parsimony is generally a virtue in predicting the causal-historical breakdown of phylogenies, it ought to be a virtue in predicting the causal-historical breakdown of phenotypes. We are justified, then, to assume that Rational Optimism provides a good rationale for inferring as few events as possible in the causal history of language evolution.¹⁰

5.3. *Parameters, Parsimony, and Plausibility*

A second supra-empirical principle of plausibility aims to reduce the number of assumptions or parameters a theory must entail. Popper (1959) believed that parameters—or, rather, a paucity of them—were an important aspect of parsimony in the philosophy of science. For him, the nature of the question was exhausted by the ductility of a theory; that is, more brittle theories—those with fewer valued parameters—are more easily falsified than very ductile ones, which can be stretched this way and that in virtue of their many manipulable parameters: “The epistemological questions which arise in connection with the concept of simplicity,” he therefore claimed, “can all be answered if we equate this concept with *degree of falsifiability*” (Popper, 1959: 140; original emphasis). Popper's is one understanding of how fewer parameters can aid a theory in achieving veridicality, though a suitably positivist one. It is prototypical of Rational Optimism, however, in that no observation or observations could diminish its force; it is properly *a priori*.

5.3.1. *The Problem of ‘Over-Fitting’*

The central idea of parameter parsimony is that fewer uncertain variables reduces the potential for error. The below figure presents a single set of data points for the two variables x and y and two polynomials which potentially describe the relationship between the points.

¹⁰ It does not follow from the likely paucity of past evolutionary events that language design is simple, or that the change which resulted in language design was simple. It does not follow because there are two circumstances under which a saltation can lead to a trait, each with different entailments for a parsimony-based metric of evolutionary plausibility.

- (i) A *minor* change in developmental chronology can lead to vast phenotypic changes. This requires only that the nature of the design in question is relatively easy to realise in physical media.
- (ii) A *major* change in developmental chronology can lead to a vast phenotypic change. This requires that all the correct developmental conditions to be in place prior to the saltation.

The state of affairs described in (i), and not that in (ii), is the scenario hypothesised in MP but it remains that cladistic parsimony licenses no inferences about the causal-history and structure of language *except that* fewer mutation-events are preferable.

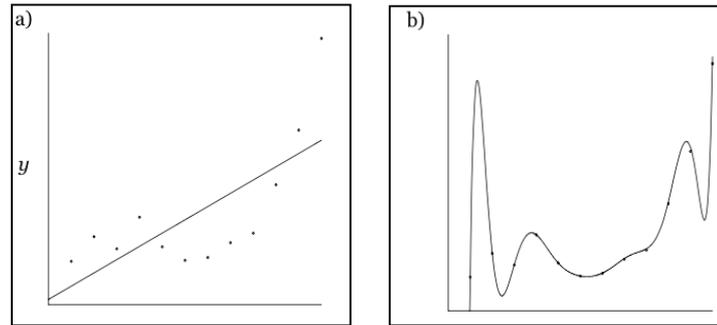


Figure 7: An illustration of a curve-fitting problem.

In (a), a simple linear regression is posited by a first-degree polynomial curve.¹¹ The nearly arbitrary straight line plot is clearly unsatisfactory, revealing nothing of interest about the relationship between x and y : A straight line can be drawn through any data and this will rarely yield any interesting analysis or further predictive accuracy.

The inverse of this point is that for any data $(x_i \dots x_n)$, there is a $n-1^{\text{th}}$ degree polynomial which plots a line perfectly through every x . We see this in (b), where every point is fitted exactly by the curve. Surprisingly, however, perfect performance on the input data will with extreme rarity translate into satisfactory predictive accuracy when the curve is extrapolated to a larger set of data. This problem affects all finite data sets (i.e., every possible data set), but is particularly troublesome for very small ones. It is known as the problem of ‘overfitting,’ where the theory incorporates experimental error and other forms of noise—such as sampling error—thus leading to an amplification of minor fluctuations in the data not relevant to the target phenomenon.

5.3.2. The Akaike Information Criterion

The Akaike Information Criterion is a method for predicting the degree of overfit for any given problem of inductive extrapolation. Sober argues that the Akaike Information Criterion provides a justification for (and metric for the degree of) unification in scientific inference. The difference between true curves and curves which contain error is estimable, according to the Akaike Information Criterion, because error is proportional to the parameters which can potentially deviate from the true curve. That is, we can estimate the degree to which higher-degree polynomial curves will overfit the data if we know the rate at which error increases with each additional parameter. However, too few parameters—like a straight line—will obviously harm goodness-of-fit. The overfitting problem is thus a question of trade-off between parsimony and goodness-of-fit. The virtue of parsimony is, in this respect, inversely linked to the potential for error.¹²

¹¹ The greater the degree of polynomial, the more complex the curve its expresses will be. The first-degree polynomial in (a) expresses a straight line, a second-degree polynomial will express a parabola, and so on.

¹² Other matters of interest are how the trade-off is to be achieved and how it is to be justified. Call the true curve C_t and the one most accurate relative to the known data C_a . We can now ask how close C_t will be to C_a , or the overfit of C_a . As set out by Forster & Sober (1994), the

5.4. Summary

We may, with Sober and Forster's imprimatur, think that the Akaike Criterion "provides a ready characterization of the circumstances in which a unified model is preferable to two disunified models that cover the same domain." (Forster & Sober, 1994: 13) The Akaike Criterion therefore provides a robust rationale for the methodology of unifying technologies in MP by establishing a concrete link between the desire to minimise the number of parameters accounting for data in a theory, and to maximise the employment of existing parameters to achieve optimal coverage. Unifications are, in essence, an exercise in minimising probable error. The frequentist interpretation of parsimony is similar in that it derives its power from a supra-empirical principle. It differs, however, in that it provides a justification for the saltationalist solution to Darwin's Problem by demonstrating the intuitive virtue of evoking common causes for evolutionary characters.

6. Physical Optimism

6.1. Spontaneity, Efficiency, and Physical Optimism

The guiding rationale of Rational Optimism is that simple science is better science. A separate concern is the degree to which the biological world, and more particularly cognition, is typified by optimal design as defined by MP. Contrary to Kinsella & Marcus' (2009) findings, evidence from the 'extended synthesis' of evolutionary biology, comparative ethology, and impressive new findings from dynamic neuroscience demonstrate saltationalism and computational optimality to be highly plausible outcomes of language evolution. The core idea is that even highly complex aspects of biological design are substantially constituted by "the surprisingly ordered systems of simple underlying processes" (Reeve & Sherman, 2001: 64f.) which emerge spontaneously and are explained by simple changes in the organisation of matter. A particular subset of these "self-organising" systems—what have been called neuro-oscillations—has been implicated in the processing of phrase-level speech signals (Ding *et al.*, 2015).

This result may vindicate Minimalist hypotheses about the origins of syntactic cognition: In light of these findings, it is highly plausible that the salient

Akaike Information Criterion provides a method for estimating the overfit of C_a with respect to the number of variables in the polynomial expressing the curve. It does so by generalising to the *family* of curves to which C_t and C_a belong, respectively, rather than considering the specific curves themselves. Call the family of curves to which C_a belongs F_a and the likelihood (in the technical sense) of the data given this family of curves $L(F_a)$. Akaike's Criterion states that the difference of C_a and C_t will be approximately equal to:

$$(i) L(F_a)_{SS} + 2k(\sigma^2)$$

In (i), k is the number of parameters in the polynomial expressing the family of curves, and SS —or the *sum of squares*—is a statistical method for finding the total variance from the mean (which therefore tracks goodness-of-fit). σ^2 relates to the size of the data sampled and reflects the notion that overfit is linked to sampling error. Notice, then, that in the absence of error ($\sigma^2 = 0$) the difference of C_a and C_t will just be the likelihood of F_a subject to SS .

aspects of language design emerged via what Benítez-Burraco (2014) has described as a perturbation of the robust equilibrium of pre-anatomically modern human's brain oscillatory rhythms. The emergence of human language can be seen through this lens as a perturbation of a highly conserved (evolutionarily ancient) self-organising system and a subsequent 'tuning' of the resulting system to result in a novel and robust phenotype. This is an appealing elaboration on the Minimalist story that provides "a better view of the genetic underpinnings of language and the molecular mechanisms that channel variation at all levels of analysis" (Benítez-Burraco, 2014: 1).

6.2. *Spontaneity, Invariance, and Darwin's Problem*

The short span of evolutionary time available to account for linguistic knowledge requires not only that there are few evolutionary events responsible for language, but also that there is a possible alteration in the organisation of physical (brain-) matter capable of producing such a phenotype in only a few steps. This scenario becomes far more plausible if there are organisations of matter which do not just reach new states rapidly, but which are also 'canalised' insofar that they will reach the required end state from any of a wide range of initial states. *Spontaneity* and *invariance* are thus key desiderata of Physical Optimism with respect to Causal-Historical Optimism—and consequently a solution to Darwin's Problem. Self-organising systems satisfy both of these desiderata; they emerge quickly and across a variety of environments. That is, the structure of some highly abstract organisations of physical matter are such that they will inexorably trend towards a state and then remain in that state indefinitely. Kauffman (1991) describes such stasis points as 'attractors' for this quality of inevitability. Stasis points are extremely robust in that they attain in a wide range of physical realisations, they emerge rapidly due to their 'attracting' capacity.¹³

Kauffman provides us with a simple example of self-organisation, which will get us on our way. "The approach begins," Kauffman starts, "by idealizing the behavior of each element in [a] system [...] as a simple binary (on or off) variable" (Kauffman, 1991: 64). That is, we ignore all but the details necessary for the general design. The particulars of this system, a network of three communicating elements, are represented in Figure 8.

¹³ These kinds of explanations fall quite naturally out of a very attractive potential framework for explaining the emergence and physical realisation of language, an approach amenable to a style of evolutionary explanation dubbed 'rational morphology' by Kauffman. This approach follows a rich tradition of biological enquiry tracing its intellectual prehistory backwards from Turing's (1952) analyses of morphogenesis, through Thompson's (1917) laws of growth, back to the original rational morphologists who counted among their numbers Goethe and Cuvier.

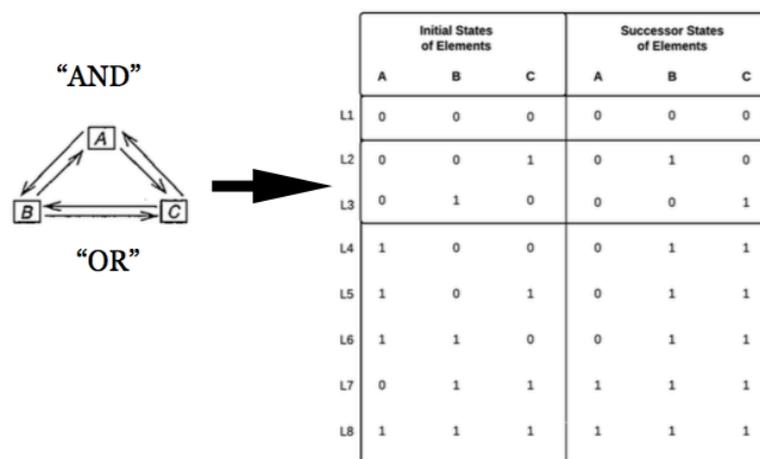


Figure 8: A self-organising network. A simple Boolean network (left) and table of possible states (right). Adapted from (Kauffman, 1991: 66).

The figure on the left is a network of three elements, each conforming to a Boolean operator, and each interacting with the other two by sending and receiving signals reflecting their current state (active '1' or inactive '0'). In this network, A functions as an AND operator, while B and C both function as OR operators. When *both* B and C are active, A will either remain or become active itself, depending on whether it was active previously. B and C will remain or become active if *either* of the other two elements are active. The table on the right describes all the ($2^3 =$) 8 starting permutations of the network and their respective successor states.

The important facts to notice are that in line L1, where all the states are inactive, there is stasis. In lines two and three, where only one of either B or C is on, the network will cycle endlessly between those two states. In all other initial states (lines four to eight), A, B, and C, will all rapidly become active and the network will again be in stasis. A remarkable upshot of these new discoveries in the area of neuro-oscillations is that both of the empirically motivated desiderata of MP—Causal-Historical Optimism and Physical Optimism—are satisfied. The framework is, furthermore, an intuitive explanation for why human nature language syntax has been so amenable to formalistic, axiom-based explanation. Self-organising systems are 'emergent', meaning they arise when highly abstract patterns of interacting matter result in a what Wagner (1989) calls an 'epigenetic trap': A robust equilibrium that is both attractive—matter in other states tends towards the equilibrium state—and invariant—matter of numerous scales is susceptible to the patten.

Consider, for instance, that

1. Syntax is *indivisible*; there is no 'half unboundedness.'
2. Syntax has the characteristic of being *discrete* in the sense that symbols and contrastive features are interpreted as independent units.
3. Syntax is readily describable in *geometric* terms, suggesting that

there is something metaphysically necessary determining the structure of syntactic cognition.¹⁴

4. Syntax exhibits the *scale-invariance* which is the most conspicuous feature of self-organising systems.

These four characteristics are conspicuously “unbiological” (Block, 1995) and are strong reasons for suspecting that self-organisation is an appropriate form of evolutionary explanation for human language.

6.3. *Homeostatic Rhythms and Cortical Entrainment*

Human natural language requires a form of hierarchical processing, which it has been hypothesised involves the Merging of syntactic objects of increasing size. This sort of scale invariance is a distinctive feature of self-organising systems, and lends itself naturally to the dynamical interpretation. That sanguinity has been known for decades: Conjecture about Fibonacci sequences is more or less *de rigueur* in considerations of evolution of language. The other reason for favouring a dynamical interpretation is more recent and inspires considerably more confidence: EEG imaging has begun to provide strong evidence that the brain comes pre-equipped with a means for encoding multiply scalar dependencies. The basis of this progress is a deepened understanding of how *homeostatic rhythms* respond to input signals. The rhythms in question are the commonplace wave frequencies—beta, delta, theta, etc.—which emerge from the excitation and discharge of cortical structures. What is novel is the discovery of how interference patterns among these frequencies encode information. Patterns interfere with one another in much the same way as people do: The loudest ones cause the most disruption.

Another way of thinking about interference is to consider the waves created by displaced water from a pebble or the stern of a boat. Waves of greater magnitude—from heavier pebbles or faster boats—will consume ones of lesser magnitude. The same is true for brainwaves. A ‘louder’ wave with greater amplitude influences ‘quieter’ ones. This becomes of great significance when the relationship between wave amplitude A and frequency f is plotted on a log scale. The result is a neat perfect line: A covaries almost perfectly with $1/f^n$. Neuroscientist György Buzsáki elaborates on why we should think this an important correlation:

[T]he inverse relationship between frequency and its power is an indication that there is a temporal relationship between frequencies: perturbations of slow frequencies cause a cascade of energy dissipation at all frequency scales. *One may speculate that these interference dynamics are the essence of the global temporal organization of the cortex.* (Buzsáki, 2006: 119; emphasis mine)

“Thus”, he claims a few pages later, “it should not come as a surprise that power (loudness) fluctuations of brain-generated and perceived sounds, like music and

¹⁴ This geometric character is particularly evident, for instance, in Kayne’s (1981) discussion of ‘unambiguous paths’ in the binding of anaphora.

speech, and numerous other time-related behaviors exhibit $1/f$ power spectra" (Buzsáki, 2006: 123).¹⁵

There has been good confirmation of the hypothesis that cortical entrainment of theta band oscillation responds to linguistically relevant syllabic units, with phase patterns observed to discriminate between actual and non-actual human natural language sentences (Ding et al., 2015). Poeppel's lab has extended this significance to the phrasal level via precisely the mechanisms of rhythmic entrainment just described (Figure 9), showing that cortical responses closely track the temporal envelopes of phrase-level syntactic objects (Ding et al., 2015: 4). The interaction of different frequencies at varying spatio-temporal scales depicted in the figure allows for hierarchical structure in signal processing.

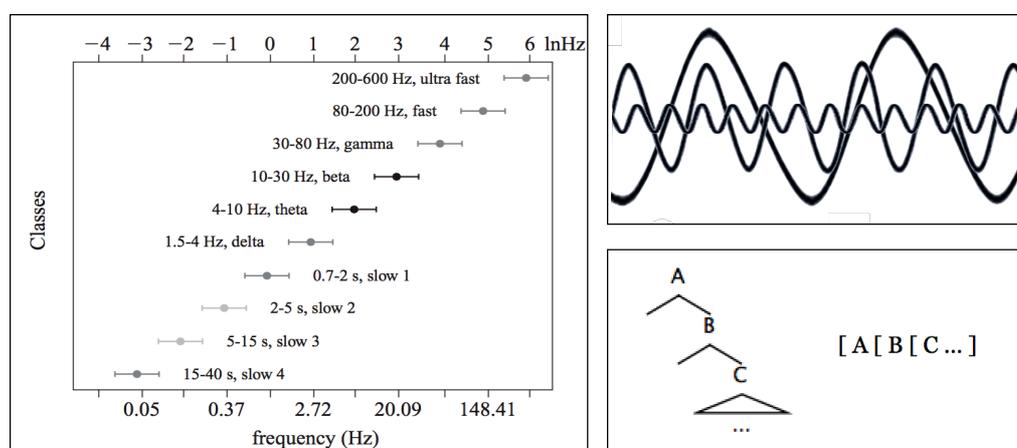


Figure 9: Cortical entrainment of temporal envelopes. The table on the left depicts ten distinct oscillating frequencies in the mammalian brain (Buzsáki, 2006: 114). Top right is an illustration of low frequency delta waves overlaid by higher frequency theta and beta waves. The interaction of these different frequencies at varying spatio-temporal scales allows for hierarchical structure in signal processing (bottom right).

What this suggests is that one *sui generis* property of human syntax—its capacity for hierarchical embedding—is a consequence of the power law holding between different rates of cortical oscillation.

These findings have recently been developed into concrete proposals for the recent evolutionary history of human syntactic cognition by Murphy (2015, 2016a, 2016b) and Ramírez (2015) which provide a plausible explanation for several syntactic phenomena (Murphy, 2015). Murphy (2016a) describes how the coupling of higher frequency gamma and lower frequency theta waves could provide a kind of “binding memory” that preserves the complex wholes of phrases.

¹⁵ This gradient has been known since the mid-nineteenth century. For instance, Weber's Law—named for Ernst Weber (*fl.* 1830–40)—noted the basic configuration in the exponential ratio of ‘just noticeable’ perceptual characters to the strength of stimulus. Well-noted examples include the phenomenological experience of heaviness compared to an object's actual weight, and the perceived versus actual change in illumination of a light source.

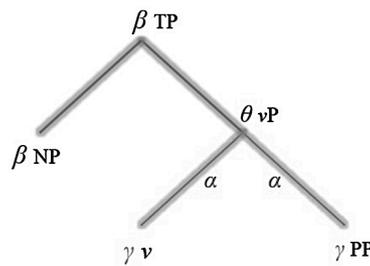


Figure 10: An 'oscillatory tree' demonstrating the alignment of syntactic-level phrases with oscillatory frequencies (Murphy, 2016a).

An interesting corollary of this schema is that it may explain *XX (Boeckx, 2013) and *{*t*,*t*} (Narita, 2015; in Murphy, 2015: 13), violations in which elements of the same category (e.g., NP, VP, CP) cannot occur adjacently.

- (12) * John_{NP} Mary_{NP} ate apples.
- (13) * [which picture of the wall]_{*i*} do you think that [the cause of the riot]_{*j*} was {*t_i*,*t_j*}?

These patterns may occur, Murphy contends, because only a single binding from the high frequency gamma wave can be sustained at one time, adding further explanatory weight to the oscillatory framework.

6.4. Efficiency and Energy-Minimisation

What, though, could possibly justify the assumption of efficiency in linguistic computation, as required by Shortest Move? Moreover, even if such a rationale exists, *why* make the assumption that it is the case for language? An oft-discussed case of actual 'in-the-world' efficiency is that of Cherniak's neural optimisation research. A good place to start is the irregularities which prompt Cherniak's interest. There are two: First, the quantity and internal angle of neuron 'arbors'—the branchings of dendritic cells (see Figure 11)—display a pattern characteristic of a diverse many natural systems—rivers, crystals, trees (actual ones, bark, leaves, etc.), *inter alia*. Second, neural components of numerous scale are organised so as to minimise the length of 'wire' (neural connective tissue) required for their interconnection. Each of these discoveries exhibits an unusual degree of *optimisation*, where optimisation is intended to denote a measure of efficiency rather than functionality. The first yields a 'local' form of optimisation, in that arbors are optimal with respect to properties of individual cells. The second is a 'global' form of optimisation which pertains to the whole network under consideration.

The two distinct kinds of optimisation have different relevance. With respect to local optimisation, our primary interest is in the mechanism of optimisation. The optimality in question is represented in Figure 11.

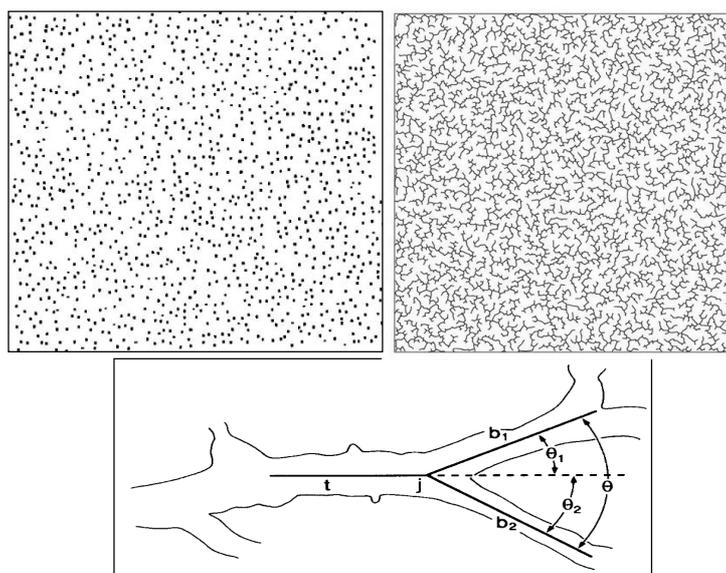


Figure 11: Steiner tree problem and branching neural axon. See text for description. Bottom adapted from (Cherniak et al., 1999: 6003).

Above are an unsolved (top left) and a solved (top right) ‘Steiner tree’—a method of calculating the minimum distance (line length) required to connect a distribution of points. This pattern is evidenced in a number of natural domains in addition to neurons—blood vessels, lung bronchi, plant roots, coral formations, antlers, rivers junctions, geological cracks, and lightning discharge patterns (Cherniak, 1992: 504). Below is an illustration of an actual dendritic arbor. The value of each internal angle θ_i and the number of branching axons b_n is observed to be close to the optimal predicted by an appropriate Steiner tree.¹⁶ This, and the aforementioned examples, are all likely to be products of a simple ‘tug of war’ energy-minimization mechanism, similar to the formation of soap bubbles and snowflakes. In all these instances, competing pressures (opponents in the tug of war) fall into an equilibrium state with minimally expensive arc angles and quantity. The significance of this mechanism is its easy congruence with the notion of self-organisation given above; efficiency and self-organisation are strange but happy bedfellows.

A second notion of optimality makes plain the relation to spontaneous order. The basic idea is similar the first, but now the metric of interest is *component placement*: We can predict with surprising accuracy the organisation of (1) the brain relative to the body, (2) the functional regions of the brain relative to one another, and (3) the internal structure of functional components like nerve ganglia. This remarkably general coverage can be achieved by invoking a single, simple rule:

¹⁶ In fact, it is not close unless it is assumed—plausibly—that the task is to conserve the *volume* of connective tissue rather than length, and that the diameter of branches is less than that of trunks. Branches will consequently have a lower ‘cost’ per unit of length compared with trunks.

The adjacency rule: If two components a and b are connected, then a and b are adjacent.

“The rule is a powerful predictor of the anatomy”, he claims, “a kind of ‘plate tectonics of the cortex’” (Cherniak, 1994a: 98). It predicts, for instance, that (a) and not (b) in Figure 12 will be the observed layout of three components. The most intuitive demonstration of the rule is the morphologically ubiquitous location of the brain in the head, a fact Cherniak claims extends naturally from the surfeit of sensorimotor connections in the morphospace’s anterior instead of its posterior.¹⁷

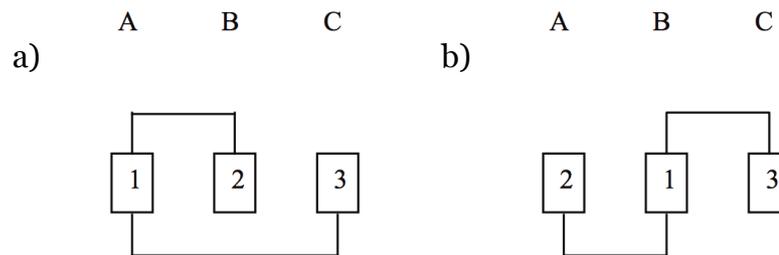


Figure 12: Representation of a component placement problem. (a) requires greater wire-length than (b), where component placement is optimal with respect to the adjacency rule (Cherniak, 1994a: 96).

Cherniak (1994a: 101) claims that “[a]n Occam’s Razor of the nervous system, the simple *logos* ‘Save wire’ invokes a significant portion of the vast neuro-wiring diagram”. It is fair to say, then, that this is no coincidence. Despite the extraordinary productiveness of the ‘save wire’ principle, neither Cherniak nor anyone else has a precise grip on what the perfect optimisation of component placement would be. This lack of understanding is not for lack of a conceptual appreciation of the task, but because of its intrinsic computational complexity: Searches for optimal paths are prototypically NP-complete. This familiar refrain throws new light on the problem of component placement:

To convey a sense of the computational intractability of exhaustive search for exact solution... it can be noted that the number of possible layouts of n components on n discrete positions (whether they form a one, two, or three-dimensional array) is $n!$ For merely the layout problem of the 50 main areas of the human cerebral cortex, there are $50! = 3.04 \times 10^{64}$ alternative placement possibilities. The number of attoseconds (10^{-18} sec) in the 20 billion year history of the universe is 10^{35} . Hence, if natural selection could test one layout per attosecond, all the time since dawn of the Universe, much less since emergence of life on Earth, would not suffice for this exhaustive search.

(Cherniak, 1994b: 2426)

¹⁷ We may wonder just how unusual the degree of observed optimisation is and consequently whether it could have been a product of mere chance. With respect to the global measure of optimisation, Cherniak estimates the null hypothesis of random component placement is improbable to a degree of certainty greater than $p = 0.0001$.

The optimality of component placement is the inverse of the '747 in a hurricane' dilemma: We are forced by necessity into the assumption that nature has employed a means of spontaneous order.

6.5. Summary

These conclusions are, inevitably, speculative; inevitably because the very idea of evolutionary plausibility pushes at the boundaries of contemporary enquiry. Yet, it is uncontroversial in most scientific domains that parsimony is one of the desiderata which can be used to determine which is preferable of two or more competing theories at a given level of organisation. The Principle of the Common Cause is sufficient to warrant the inference of parsimony with respect to the *number of causes* responsible for language design. A distinct motivation, but one no less important, is that the brittleness of a theory—its paucity of parameters for potential error—motivates a unification-based approach. These conceptions of 'Rational' optimism apply not to a theory of causes implicated in the design of syntax, but, rather, to the theory of syntax which is the target of that explanation. Physical Optimism follows naturally from the characterisation of language as self-organising, and goes part of the way towards explaining how an *independently motivated* efficiency condition may be realised in physical media which we suspect is self-organising. The presumption of Physical Optimism also solves Darwin's Problem by providing a plausible scenario in which spontaneous emergence of order can overcome underdetermination.

References

- Baltin, Mark R. 2011. The copy theory of movement and the binding-theoretical status of A-traces: You can't get there from here. *NYU Working Papers in Linguistics* 3, 1–28.
- Benítez-Burraco, Antonio. 2014. Biological noise and H2A.Z: A promising connection for language. *Frontiers in Genetics* 5: 463, doi: 10.3389/fgene.2014.00463.
- Berwick, Robert, Kazuo Okanoya, Gabriël Beckers & Johan Bolhuis. 2011. Songs to syntax: The linguistics of birdsong. *Trends in Cognitive Sciences* 15, 113–121.
- Boeckx, Cedric. 2009. The nature of merge: Consequences for language, mind, and biology. In: Massimo Piattelli-Palmarini, Juan Uriagereka & Pello Salaburu (eds.), *Of Minds and Language: A Dialogue with Noam Chomsky in the Basque Country*, 44–57. Oxford: Oxford University Press.
- Boeckx, Cedric. 2013. Merge: Biolinguistic considerations. *English Linguistics* 30, 463–484.
- Block, Ned. 1995. The mind as the software of the brain. In Edward E. Smith & Daniel Osherson (eds.), *An Invitation to Cognitive Science*, vol. 3: *Thinking*, 2nd edn., 377–425. Cambridge, MA: MIT Press.

- Boskovic, Željko & Jairo Nunes. 2007. The copy theory of movement: A view from PF. In Norbert Corver & Jairo Nunes (eds.), *The Copy Theory of Movement*, 13–74. Amsterdam: John Benjamins.
- Buzsáki, György. 2006. *Rhythms of the Brain*. Oxford: Oxford University Press.
- Cherniak, Christopher. 1992. Local optimization of neuron arbors. *Biological Cybernetics* 66, 503–510.
- Cherniak, Christopher. 1994a. Philosophy and computational neuroanatomy. *Philosophical Studies* 73(2–3), 89–107.
- Cherniak, Christopher. 1994b. Component placement optimization in the brain. *The Journal of Neuroscience* 14(4), 2418–2427.
- Cherniak, Christopher, Mark Changizi & Dong-Wha Kang. 1999. Large-scale optimization of neuron arbors. *Physical Review E* 59(5), 6001–6009.
- Chomsky, Noam. 1956. Three models for the description of language. *Information Theory* 2(3), 113–124.
- Chomsky, Noam. 1959. On certain formal properties of grammars. *Information and Control* 2(2), 137–167.
- Chomsky, Noam. 1981. *Lectures on Government and Binding*. Dordrecht: Foris.
- Chomsky, Noam. 2000a. *New Horizons in the Study of Language and Mind*. Cambridge: Cambridge University Press.
- Chomsky, Noam. 2000b. Minimalist inquiries: The framework. In Roger Martin, David Michaels & Juan Uriagereka (eds.), *Step by Step: Essays on Minimalist Syntax in Honor of Howard Lasnik*, 89–155. Cambridge, MA: MIT Press.
- Chomsky, Noam. 2005. Three factors in language design. *Linguistic Inquiry* 36, 1–22.
- Chomsky, Noam. 2007. Approaching UG from below. In Uli Sauerland & Hans-Martin Gärtner (eds.), *Interfaces + Recursion = Language? Chomsky's Minimalism and the View from Syntax–Semantics*, 1–29. Berlin: Walter de Gruyter.
- Ding, Nai, Lucia Melloni, Hang Zhang, Xing Tian & David Poeppel. 2015. Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience* 19, 158–164.
- Forster, Malcolm & Elliott Sober. 1994. How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *The British Journal for the Philosophy of Science* 45, 1–35.
- Hauser, Marc D., Noam Chomsky & W. Tecumseh Fitch. 2002. The faculty of language: What is it, who has it, and how did it evolve? *Science* 298(5598), 1569–1579.
- Higginbotham, James. 1984. English is not a context-free language. *Linguistic Inquiry* 15, 225–234.
- Hornstein, Norbert. 2001. *Move! A Minimalist Theory of Construal*. Oxford: Wiley Blackwell.
- Kauffman, Stuart. 1991. Antichaos and adaptation. *Scientific American* 265, 64–70.
- Kayne, Richard S. 1981. Unambiguous paths. In Robert May & Jan Koster (eds.), *Levels of Syntactic Representation*, 143–183. Dordrecht: Reidel. [Reprinted in Richard S. Kayne, *Connectedness and Binary Branching*, 129–163. Dordrecht: Foris, 1984.]
- Kinsella, Anna R. 2009. *Language Evolution and Syntactic Theory*. Cambridge: Cambridge University Press.

- Kinsella, Anna R. & Gary Marcus. 2009. Evolution, perfection, and theories of language. *Biolinguistics* 3(2-3), 186-212.
- Marcus, Gary. 2009. *Kluge: The Haphazard Evolution of the Human Mind*. Boston, MA: Houghton Mifflin Harcourt.
- Murphy, Elliot. 2015. The brain dynamics of linguistic computation. *Frontiers in Psychology* 6: 1515, doi: 10.3389/fpsyg.2015.01515.
- Murphy, Elliot. 2016a. Evolutionary monkey oscillomics: Generating linking hypotheses from preserved brain rhythms. *Theoretical Linguistics* 42(1-2), 117-137.
- Murphy, Elliot. 2016b. The human oscillome and its explanatory potential. *Biolinguistics* 10, 6-20.
- Narita, Hiroki. 2015. *{t,t}. In Ulrike Steindl, Thomas Borer, Huilin Fang, Alfredo García Pardo, Peter Guekguezian, Brian Hsu, Charlie O'Hara & Iris Chuoying Ouyang (eds.), *WCCFL 32: Proceedings of the 32nd West Coast Conference on Formal Linguistics*, 286-295. Somerville, MA: Cascadilla Proceedings Project.
- Parker, Anna R. 2006. Evolution as a constraint on theories of syntax: The case against minimalism. Edinburgh: University of Edinburgh dissertation.
- Popper, Karl. 1959. *The Logic of Scientific Discovery*. London: Hutchinson.
- Ramírez, Javier. 2015. Locality in language and locality in brain oscillatory structures. *Biolinguistics* 9, 74-95.
- Reeve, Hudson K., & Philip Sherman. 2001. Optimality and phylogeny: A critique of current thought. In Elliott Sober & Steven Orzack (eds.), *Adaptationism and Optimality*, 64-113. Cambridge: Cambridge University Press.
- Reichenbach, Hans. 1956. *The Direction of Time*. Berkeley, CA: University of California Press.
- Simon, Herbert. 1969. *The Sciences of the Artificial*. Cambridge, MA: MIT Press.
- Sober, Elliott. 1988. *Reconstructing the Past: Parsimony, Evolution, and Inference*. Cambridge, MA: MIT Press.
- Sober, Elliott. 2015. *Ockham's Razors*. Cambridge: Cambridge University Press.
- Thompson, D'Arcy Wentworth. 1917. *On Growth and Form*. Cambridge: Cambridge University Press.
- Turing, Alan. 1952. The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 237(641), 37-72.
- Wagner, Günter. 1989. The origin of morphological characters and the biological basis of homology. *Evolution* 43(6), 1157-1171.

Michael R. Levot
University of New South Wales
Department of Humanities and Languages
Morven Brown Building, UNSW
Sydney, NSW 2052
Australia
m.levot@unsw.edu.au