Synthese
https://doi.org/10.1007/s11229-019-02290-x

S.I.: EXPLANATIONS IN COGNITIVE SCIENCE: UNIFICATION VS PLURALISM

Linguistics and the explanatory economy

Gabe Dupre[1]

Received: 12 July 2018 / Accepted: 11 June 2019

Abstract
I present a novel, collaborative, methodology for linguistics: what I call the 'explanatory economy'. According to this picture, multiple models/theories are evaluated based on the extent to which they complement one another with respect to data coverage. I show how this model can resolve a long-standing worry about the methodology of generative linguistics: that by creating too much distance between data and theory, the empirical credentials of this research program are tarnished. I provide justifications of such methodologically central distinctions as the competence/performance and core/periphery distinction, and then show how we can understand the push for simplicity in the history of generative grammar in this light.

Keywords Philosophy of linguistics · Philosophy of cognitive science · Generative syntax · Philosophy of science · Philosophy of language

1 Introduction

It is a truism that if linguistics is to be a genuine empirical science, then its proposals must be sensitive to the observational data. Strikingly, theorists of many different stripes have claimed that generative linguistics, the most well known and influential research program in linguistic theory, does not meet this requirement. To see why, consider the following exchange:

A: I saw Maria proof-reading her draft.
B: What is she writing?

✉ Gabe Dupre
Gdupre@humnet.ucla.edu

[1] University of California, Los Angeles, Los Angeles, USA

Published online: 20 June 2019

A: *(A loud motorcycle drives by, making A's response hard to discern.)* She is writing a screenplay.
B: She is writing what?

Standard generativist theories predict that B's first question, in which the wh-expression 'what' has been raised from the position at which it receives its semantic interpretation (as the object of 'writing') to the beginning of the sentence, is grammatical. However, they also predict, for reasons to be discussed in detail later, that B's second question is ungrammatical. The displacement of 'what' observed in the former question is claimed to be *mandatory*, so that sentences in which such movement does not occur are deemed ungrammatical.

This seems like a refutation of the theory. A grammar for English is supposed to determine, for each expression, whether it is well-formed in English or not. Intuitively, both of B's questions are well-formed, but standard generative theories predict that only one of them is. One might think that this would lead generative grammarians to reformulate their theories so as to no longer make this bad prediction. But this is not what they typically do. Instead, they often simply exclude such data from the purview of their theory. Various tools in the methodological arsenal of the generativist seem to serve exactly this function: the competence/performance, acceptability/grammaticality, and core/periphery distinctions, to be discussed at length later, all function to distinguish between data that are relevant to (dis-)confirming the relevant hypotheses and data that are not.

This approach, in the eyes of many, undermines the status of generative linguistics as an empirical science. The ability of generativists to select, often *post hoc*, which data are relevant to their theories, has been claimed to be unscientific. According to this objection, by adopting such strategies the generativist is able to immunize her theory from the data, rendering it unfalsifiable.

In this paper, I shall propose a novel account of the methodology of linguistics, involving what I call the 'explanatory economy'. This methodology should make clear why the objections to the generativist program miss the mark. This picture focuses on the ways in which, prior to inquiry, it is not known which phenomena are to be explained by which theories. The impression that some datum is within the purview of a particular theory is revisable, subject to a more developed understanding of the theory in question and other nearby theories. In making such a revision, and excluding this datum from the scope of one's theory, a theorist is relying on this datum being explicable by some other non-competing theory. I describe such a theorist, and by extension her theory, as incurring an 'explanatory debt'. This debt is discharged when some other non-competing theory indeed provides such an explanation. If this debt is unlikely to be paid, this apparent counterexample may be re-evaluated as indeed within the scope of the original theory and thus may count as a genuine counterexample. In this way, the success of a theory is dependent on the success of the other theories to which it is indebted.

The focus on this dependency differentiates my approach from traditional approaches to confirmation. While these approaches, from Hempel and Oppenheim's (1948) Hypothetico-Deductivist model, Popper's (1959/2002) falsificationism and

through to more recent Bayesian approaches,[1] have produced illuminating pictures of the relationship between data and hypothesis, they have all focused on the confirmation accrued to a particular theory in isolation from other proposals. These approaches are *individualistic* in that the confirmation of one theory is independent of the explanatory power of all other theories.[2] My approach differs in that the extent to which one theory is confirmed itself depends on the explanatory power of distinct theories. In particular, whether certain data (dis-)confirm a particular theory will depend on whether these data can be accounted for by other theories consistent with the former. I will call my approach *collectivist* in that it views theories as confirmed based on how well they collectively, rather than individually, account for the observations.

On this approach, then, confirmation is a function of both the explanatory/predictive coverage of a theory *and* the degree to which the debts it has incurred are likely to be discharged. Consider two rival theories, A and B, which provide predictions/explanations of overlapping but distinct data sets of roughly equal size and import. From an individualistic perspective, there will be little to choose between these two theories.[3] However, this may not be so from a collectivist standpoint. The explanatory economy may enable us to choose between these theories based on the way these theories square with others. In particular, if the debts accrued to theory A, but not B, are plausibly discharged by some third theory C (i.e. C can explain (away) apparent counter-examples to A, but not B), then A is better confirmed than B despite having roughly similar predictive/explanatory power.

Such a methodology is crucial for linguistic theory, as the observations (typically, the linguistic judgements of a native speaker) are products of a wide range of cognitive systems. Specifically linguistic capacities play an important role, but due to the importance of memory, sensory-motor systems, speaker intentions, etc., these capacities are not directly reflected by the data. The explanatory economy provides a way to factor out these influences: our theories of specifically linguistic capacities are committed to explaining those observations which cannot be accounted for by theories of other systems. This approach fits particularly nicely with, but does not presuppose, a modular theory of mind. On a modular picture, language constitutes just one of many relatively independent cognitive systems. As linguistic theory develops, linguists and other theorists make guesses about which data are reflective of which systems. The explanatory economy is a tally of which theorists have made which guesses, and the extent to which such guesses are successful.

One upshot of the explanatory economy approach is that it can enable us to reinterpret some important debates. In particular, the individualistic view of theory

[1] The *locus classicus* here being Earman (1992).

[2] This is not to say that these approaches *ignore* alternative theories, just that each theory's confirmation is determined independently. Discussions of the confirmation of distinct theories within these approaches are comparative. For example, one of the most fruitful applications of Bayesian approaches to confirmation is Bayesian Model Selection, the use of Bayesian statistical tools to select between rival models of observed data (See e.g. Wasserman (2000) for an overview). This approach involves looking at multiple models, but it does so by independently assessing each, and then comparing these assessments, rather than treating the confirmation of one theory as itself dependent on the successes of another.

[3] Let us assume that A and B are likewise roughly equivalent with respect to their extra-empirical virtues (simplicity, elegance, etc.).

confirmation suggests that alternative approaches to linguistics should be evaluated on a winner-takes-all basis. Generative linguistics and some of the popular alternative accounts, such as usage-based grammar, are treated in this way: either linguistic competence is as described by the nativist generative theory, or by the empiricist usage-based theory. I shall argue instead that these approaches may be better treated as complementary. Neither can account for all of the phenomena, but this is because they each target different aspects of our linguistic capacities. This collectivist approach enables us to get the best of both worlds: universal properties of the linguistic faculty, as described by generative theories, can account for similarities across language and across speakers, while lexical and peripheral idiosyncrasies can be explained by empiricist theories. While this conciliatory approach may not vindicate the most ambitious claims of either approach, hopefully it will enable both to be better confirmed in their proprietary domains.

My explanatory economy approach bears a resemblance to the line of work focusing on the 'division of cognitive labor'.[4] At a certain relatively abstract level, my approach and that found in this tradition are aimed at answering similar questions about the structure of scientific practice: how to determine which projects to investigate, and how the investigation of one area influences the investigation of others. However, at a more fine-grained level, these approaches are quite different in their aims. In particular, traditional work on the division of cognitive labor has focused on the questions of how the behavior of individual researchers does and should depend on that of other researchers, whereas my interest is in the ways in which particular hypotheses or theories are legitimized by work done in other areas. In fact, I take the phenomena I shall describe to present difficulties for the modeling work produced in this tradition. I shall spell out these similarities and differences in more detail in Sect. 10, distinguishing between work which aims to uncover science's *explanatory structure* and that which aims to uncover its *institutional structure*.

2 Competence and performance

At a very general level, confirmation of linguistic proposals involves comparing the outputs of proposed generative systems (i.e. sets of rules and constraints which determine the ways in which properties of simple expressions determine properties of the complex expressions into which they are combined) with linguistic judgements and behavior. For example, if a sentence S of a natural language L is judged by native speakers to be exactly three-ways ambiguous, then a proposed generative system for L which allows for the generation of exactly three structures pronounced as S, but with distinct interpretations, will be thereby (partially) confirmed. A system which produced exactly two or four such structures would be thereby disconfirmed.

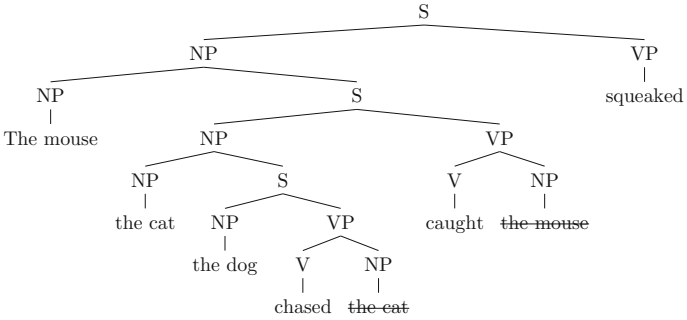
However, in many cases, this simple approach to confirmation fails to capture the practice of generative linguistics. Structures generated by the generative systems correspond to natural language sentences which are not deemed acceptable, and sen-

[4] This work originates with Kitcher (1990) and its development in Strevens (2003, 2006), but I am including under this banner also the model-based work of e.g. Weisberg and Muldoon (2009) and Zollman (2007).

tences deemed acceptable fail to correspond to structures generated by these systems.[5] Despite this, the generative theory is not thereby regarded as falsified. For example, consider:

(1) The mouse the cat the dog chased caught squeaked.
(2) Tigers tigers tigers fight fight fight.

These sentences seem, at first hearing, to be word salad. They are judged unacceptable. However, it can be shown that they are the products of the normal application of grammatical rules that in similar cases produce perfectly normal sentences.[6] Here is a highly simplified version of the grammatical structure for sentence (1):



Note that the complexity of these structures is generated by perfectly normal grammatical rules. Starting with a sentence containing a transitive verb, we can form a complex NP headed by the object of the verb in the way diagrammed in the tree above (e.g. the sentence "the dog chased the cat" becomes the noun-phrase "the cat the dog chased"). Any syntactic theory must allow for such constructions. But once this is allowed, iteration of this process leads to unacceptable results like (1) and (2).

Linguists could, faced with examples like (1) and (2), posit more complicated grammatical rules for English (e.g. constraints on the number of possible embeddings). However, this is not what generativists do. Instead, they distinguish between grammaticality and acceptability. The distinction between grammaticality and acceptability, and more generally between competence and performance, has been methodologically central to the generativist program since at least Chomsky (1965). A sentence is acceptable when native speakers judge it so. Acceptability is thus a broadly *observational* term. However, grammaticality is a theoretical term. That is, there is no characteristic observable linguistic behavior indicative of grammaticality. A sentence is grammatical if and only if it has a syntactic structure generatable by a speaker's internalized linguistic competence. This distinction is thus an instance of the distinction between

[5] This rather clunky terminology of sentences 'corresponding to' structures and *vice versa* is motivated by the fact that the linguistic system produces hierarchical structural descriptions, which must then be linearized to be pronounced as sentences. As I will use the terms, a sentence is ambiguous when it corresponds to multiple structures.

[6] For reasons of space, I am here restricting my attention to sentences which are generated by the proposed grammar but not judged acceptable by native speakers. The opposite phenomenon, of sentences judged to be acceptable but not generated by proposed grammars, is widespread as well. Examples of this sort include 'linguistic illusions' such as "More people have been to Australia than I have" which seem at first hearing to be perfectly legitimate English sentences, but for which, upon reflection, no coherent interpretation can be provided.

competence, the structure of an underlying capacity, and *performance*, the behavioral output of interaction between this and many other cognitive systems. Acceptability is dependent, to some degree, on grammaticality, but on numerous other properties of a speaker's psychology as well. The generativist claims that their goal is to describe competence (grammaticality) and not performance (acceptability). Thus, the generativist can claim that (1) and (2) are indeed grammatical, just as their theory predicts. The unacceptability of these expressions is then claimed to reflect limitations on performance, and so does not provide a counterexample for a theory of competence.

3 Core and periphery

Alongside the competence/performance distinction, generativists have also invoked the distinction between language's core and periphery.[7] It is claimed that some aspects of the language system (the core) are essential to it, while others are more or less optional (the periphery). The core may include principles—universal constraints on natural languages—and parameters—principle-schemas allowing for highly constrained variation (typically a selection between two options). The periphery will include "borrowings, historical residues, inventions and so on, which we can hardly expect to—and indeed would not want to—incorporate within a principled theory of [Universal Grammar]." (*ibid* p. 8). The periphery, then, includes a wide variety of factors outside of the core language faculty which may play a role in influencing linguistic behavior.

Peripheral elements of the language will be relatively unsystematic and idiosyncratic in comparison to the deep (and species-universal) properties of the core. The echo-questions mentioned in the introduction provide a case in point. Consider the following paradigm:

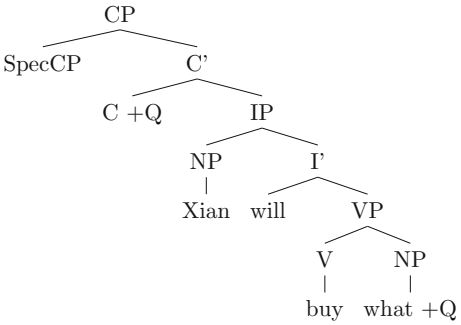
(3) Xian will buy a car.
(4) What will Xian buy?
(5) *Xian will buy what?

(4) is the interrogative form of (3), where the direct object is the element being questioned. The basic story for how questions are formed says that (4) is formed on the basis of a structure like (3) with a wh-expression in object position which is moved to the front of the sentence.[8] The question is: why in English do we ask this question with (4) rather than (5)? That is, why, given that we are questioning the object of the verb 'buy', do we pronounce the wh-word at the beginning of the sentence, rather than the end?

[7] This distinction dates back at least to Chomsky (1981) which is based on lectures given in 1979.

[8] I.e. "What$_1$ will Xian buy [what$_1$]?" I will use this convention of square brackets indicating the original location of a moved element throughout this paper.

The standard way to explain this kind of 'displacement' in the generative program is in terms of feature-driven movement.[9] Some expressions are described as having certain kinds of 'features', properties which call for a 'valuation' by other elements of the tree structures in which they are found. Without this valuation, these expressions, and the trees in which they are found, are uninterpretable. However, such valuation can only occur locally: elements of the tree that are too far away cannot satisfy this constraint. For this reason, when a tree is generated in which such features cannot be valued, it may be necessary to re-arrange some of the elements of the tree so as to enable valuation to occur. A simplified underlying structure for sentence (4), prior to such movement would be:

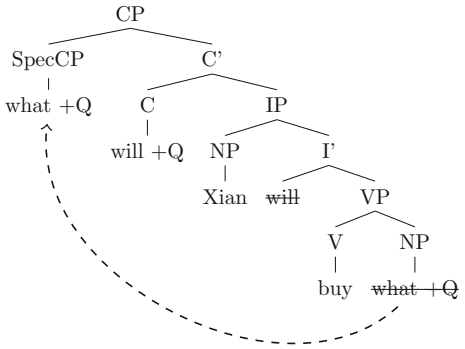


This tree structure indicates that the underlying sentential clause is "Xian will buy what." Two facts matter for our purposes. Firstly, this sentential clause (IP) is attached to an unpronounced complementizer (C). Because this is an interrogative structure, this complementizer is assumed to contain an interrogative feature (+Q), which must be valued.[10] Secondly, the wh-expression 'what' has an analogous feature, capable of valuing the +Q feature on the complementizer, but these features are originally too far apart to achieve this. For this reason, this latter expression is moved[11]:

[9] Generativist accounts which don't make use of the machinery of features give alternative explanations. One traditional picture claimed that wh-expressions are *operators*, and so may not serve as arguments of verbs. When an operator is moved, it leaves behind a *trace* which is a suitable argument. So, on this account, movement serves to ensure that each verb has the right number of arguments. For various reasons, contemporary Minimalist approaches reject this picture. For concreteness, I will focus on the feature-driven account, but everything I say should apply *mutatis mutandis* to alternative theories of displacement.

[10] Think of C here as an unpronounced analog of 'whether'. Unpronounced complementizers for declarative clauses will be analogous to 'that', and will have non-interrogative features (−Q).

[11] The expression indicating tense, 'will', is, for similar reasons, moved to the site of C. For simplicity, I shall not discuss this for the purposes of this paper.



After movement, 'What', originally generated lower-down in the tree than the subject 'Xian', is now pronounced at the beginning of the sentence. 'What' has been attached to the top of the tree, a position known as 'SpecCP' posited primarily as a host location ('landing site') for moved elements like wh-expressions. By moving this expression into this higher location, it is now close enough to the complementizer C that it can value the +Q feature, making the sentence interpretable.[12]

By making these several assumptions (all sentential clauses are attached to Cs with illocutionary force features, these features require valuation for interpretation, etc.) generative grammars are able to explain certain phenomena and account for surprising cross-linguistic regularities.[13] However, such an account has *prima facie* counterexamples. Take sentence (5) for example. According to the proposal just given, this question should be uninterpretable. The explanation for the displacement of 'what' in (4) relied on the assumption that the unvalued (+Q) feature of the sentential complementizer is uninterpretable prior to this displacement. This displacement is motivated solely to avoid this uninterpretability. But, assuming that (5) has exactly the structure of (4) prior to displacement, (5) is predicted to be uninterpretable as well. But such questions are, in the right discourse context, perfectly acceptable. These 'echo-questions' are acceptable in response to examples like (3), to indicate either that the speaker didn't hear the direct object or (perhaps) that the direct object was surprising/inappropriate.

There are several kinds of response one can give to such apparent counterexamples. The simplest is to accept that they are indeed counterexamples and thus to re-formulate the principles of our core grammatical theory so as to avoid predicting that they are unacceptable. There are, however, good reasons for not doing this. The grammatical principles that together make this seemingly bad prediction play essential roles in a wide variety of linguistic explanations. Revising these rules thus has important consequences across linguistic theory, oftentimes negative consequences. For this reason, it is often preferable to retain the theories and deny that the observed phenomena are genuine counterexamples. In order to see this, I will briefly look at some theoretical consequences of revising these principles.

Sobin (1990) and Huddleston (1994) both claim that, despite appearances, echo-questions like (5) are not necessarily interrogative, and instead inherit the C(omp)-

[12] See Chomsky (1995) for the details.

[13] See e.g. Bošković (1998) for an application of these proposals to account for differences between the grammar of English and French questions.

feature (interrogative, imperative, declarative etc.) from the utterances to which they are responsive. As sentence (3), to which sentence (5) is an appropriate response, is a declarative sentence, on this account (5) is also. This would mean that the unpronounced complementizer in sentence (5) does not have an interrogative (+Q) feature. The feature-driven account of movement I have presented explains the movement in (4) on the grounds that the +Q feature heading this sentence requires valuation. As, on this view, there is no +Q feature present in (5), this theory would thus no longer predict that 'what' must move in order to check this feature. So the wh-expression remaining in place is expected, and no bad prediction is generated.[14]

However, this response comes at a cost. If (5) is not an interrogative, then it is a counterexample to the 'wh-criterion'.[15] This criterion entails that wh-expressions (with a +Q feature) can only be found within interrogative clauses. Sentence (5), on the analysis just given, will thus be a counterexample to this, as it is claimed that this sentence is non-interrogative despite its containing a wh-expression. This principle, however, does a lot of work in grammatical theory.[16] If we reject the wh-criterion, allowing for wh-expressions in the absence of interrogative complementizers as we must if (5) is not interrogative, echo-questions are no longer problematic, but our explanations for other sorts of grammatical phenomena are undermined, including of course the explanation for why (5) is usually unacceptable.

Carnie (2013) (pp. 382–383) gives an alternative account, according to which echo-questions involve a *sui generis* complementizer, marked +INT (to indicate that wh-expressions within its complement receive a special intonation), and the wh-expressions found in such questions do not have their normal +Q feature. This +INT feature does not require valuation, and so does not motivate movement. This would avoid conflicting with the wh-criterion, as this rule applies only to wh-expressions and complementizers which have their normal interrogative features. However, this proposal suffers from a number of drawbacks. Firstly, it is clearly *ad hoc*. This com-

[14] This would not explain the failure of movement in echo-questions that are interrogative, such as B's response in the following dialogue: A:"Will Xian buy a car?" B: "Will Xian buy what?".

[15] See, for example, May (1985) (p.17) and Rizzi (1996). Note that such a principle also suggests that exclamatives ("Maria will do what?!") are syntactically interrogative.

[16] Among other things, it explains the distribution of wh-expressions within attitude reports. For example, 'believe' and 'wonder' take only declarative and interrogative complements, respectively, as indicated by the explicit complementizers they may take:

(6) Abdul believes that Xian bought a car.
(7) *Abdul believes whether/if Xian bought a car.
(8) Abdul wonders whether/if Xian bought a car.
(9) *Abdul wonders that Xian bought a car.

These attitude verbs also differ in whether the complements they take may include pronounced wh-expressions:

(10) *Abdul believes who bought a car.
(11) Abdul wonders who bought a car.

The wh-criterion provides an explanation for this pattern: 'wonder' must combine with a clause featuring a +Q complementizer, and thus this clause may include a wh-expression, while 'believe' cannot combine with such a clause, and so its complements are prohibited from including wh-expressions. Two distinct distributional facts (about complementizers and wh-expressions) can thus be subsumed under one generalization.

plementizer is introduced precisely and solely in order to account for these exceptional cases. This might be OK if there were some independent predictions that such a posit could make, but there are reasons to think this will not be the case.[17] This means that as well as positing this extra complementizer, we would need to introduce very specific constraints on when it is licensed in order to prevent over-generation. Another worry is that there is no overt version of such a complementizer in English (unlike +Q 'whether/if' and −Q 'that'), or, as far as I know, any other language. If this complementizer were present in the grammar, we might expect to see it overtly somewhere. Relatedly, it leaves unexplained why in English (and, as far as I know, every other language) these independent lexical items (wh-expressions with and without a +Q feature) are pronounced in the same way. These further facts about the strange distribution of these *sui generis* COMPs and wh-expressions must themselves be explained.

Yet another proposal is that these examples do involve the normal movement of the embedded wh-expression to the position at the beginning of the sentence (Spec-CP), but that this movement is *covert*, occurring silently. On such a proposal, the difference between normal wh-questions and echo-questions is just a matter of which copy of the wh-expression is pronounced. The underlying structure of both expressions is the same, but in the former the higher copy of the expression is pronounced, whereas in the latter the lower copy is. This would make the difference between these expressions not one of syntax, but of the way that syntactic items are pronounced (or 'spelled-out'). This account is plausible, and fits more neatly into standard Minimalist theorizing. However, it is insufficient in two ways. Firstly, it seems to amount simply to a stipulation that wh-expressions are pronounced in one location in some contexts and at other locations in others, and so it leaves the explanation for why this is the case open. It seems quite likely that such a rule will need to be explained in the empiricist ways I will describe later on, in Sect. 7. Secondly, it again provides a counter-example to compelling linguistic generalizations. In this case, the Edge Condition on Copy Deletion (Trinh 2011) states that lower copies are deleted (i.e. unpronounced) when they are the final expression of a phrase. In the case above, 'what' does end the embedded IP and so is predicted to be deleted. Indeed, it is this prediction which allows us to make sense of the usual (non-echo) wh-questions. Relaxing this rule would thus remove echo-questions as counter-examples to our theories, but at the cost that the pronunciations of normal wh-questions are no longer explained. As in the cases above then, we can provide theories which account for these observations, but only at the cost of complicating our linguistic theories to the point that we lose the ability to explain other phenomena.

Of course, these are not knock-down arguments, and there is much more that could be said about ways of accounting for echo-questions within the scope of generative grammar.[18] Going down all these paths would take me well away from the topic of

[17] It seems essential to the explanation that +INT complementizers may *only* occur in echo-questions. That echo-questions can occur only in certain discourse situations is a datum in need of explanation. If these sentences, when featuring this +INT complementizer, are perfectly grammatical, Carnie owes an explanation of why they cannot be used discourse-initially.

[18] There are also theorists who, in part due to such examples, reject mainstream generative theories entirely. Those working within the framework of Head-driven Phrase Structure Grammar (HPSG), for example, reject the claim that expressions (e.g. wh-expressions) move at all. Such theorists also typically view competence as a much broader phenomenon, including context-sensitivity and appropriateness to discourse, than do those

this paper. The point of these examples is just to show how the various explanations in syntax are entangled. Modifying the theory to account for apparent counterexamples in one domain often undermines explanations in others, and raises more questions than it purports to solve.

For this reason, it is often attractive, instead of coming up with a theory-internal account of how these apparent anomalies arise, to simply exclude these phenomena from the scope of the theory. The posits of generative grammar, such as the wh-criterion, the requirement that certain features be valued, etc., describe the essential core of our linguistic capacities. However, learned exceptions and relaxations of these core rules, such as the allowance of violations to mandatory movement rules in certain specified discourse situations, can arise within the periphery. These peripheral phenomena are thus outside of the scope of generative grammatical theory, a theory of the linguistic core, and so cannot serve as counterexamples to these theories.[19]

4 The Galilean style and its critics

The strategy exemplified by these two case studies has recently been dubbed 'The Galilean Style'. For example, Chomsky (2002) says "[Galileo] was willing to say 'Look, if the data refute the theory, the data are probably wrong'." (p. 98). The driving thought behind the Galilean style is that the complex interactions responsible for observable phenomena preclude deep and general theories from making accurate predictions of most data. Given the impossibility of achieving both explanatory depth and empirical adequacy, the aim is to capture one underlying system, which plays a role in determining the observable behavior of the complex whole. Because this system will be one of many factors contributing to the determination of observable behavior, it is to be expected that there will be many features of this behavior that are not predictable from an understanding of this particular system alone.[20]

The Galilean approach relies on both metaphysical and methodological assumptions. The metaphysical assumption is that the underlying systems are relatively simple. But what is observed is almost always an interaction effect of the operation of many different systems and so it is typically very difficult to make good inferences from observed data to its causes (and *vice versa* predictions from causes to effects). This is clearest in the case of language, where produced utterances depend on the interaction

in mainstream generative grammar. This means that the issue of echo-questions is viewed very differently. See Ginzburg and Sag (2000) for a classic discussion, and Purver (2004) for a computational approach along these lines. Discussing the viability of this alternative strategy would take me too far away from my goal of elucidating a move typical of mainstream generative theory.

[19] I believe a wide range of more-or-less systematic linguistic phenomena fall into this class. One example is certain semi-idiomatic uses of reflexives which occur without antecedents, as in "Politicians, such as yourself, ruin everything" or "I hope to see Leela and yourself at the party." Labov (1975) (p.107) describes another example of American dialects which violate purported constraints on the distribution of 'any'. Zwicky and Pullum (1987) suggest a view akin to the one I am suggesting for dealing with unusual grammatical phenomena in expressive speech.

[20] Note that this is not equivalent to saying that the laws posited by generative linguistics hold only *ceteris paribus*. The laws may themselves be perfectly strict, in their application to this underlying system. The observed behavior is not covered by these laws, and thus cannot provide counterexamples.

of purely linguistic capacities (syntax, phonology, semantics), broader psychological phenomena (memory, processing heuristics), speaker intentions, etc. For this reason, generativist methodology proceeds largely by ignoring most of the observable data and looking for phenomena which shed light more directly on the underlying capacities/systems of interest. On the assumption that the observable data differentially reflect the workings of different underlying mechanisms, with some being full-blown interaction effects but others being indicative of the outputs of particular systems, it is thought that judicious selection of relevant data can shed light on the underlying system of interest, the human language faculty, even though the influence of this system on the observations is highly indirect in many cases.

The generativist tradition has long been moving in this direction, towards simple descriptions of underlying systems. This has culminated in the *Strong Minimalist Thesis* (Chomsky 1995) that language is an optimal solution to the problem of mapping signs onto meanings. That is, the grammatical principles of the human language faculty provide the computationally most efficient means of mapping phonological representations onto semantic representations. In particular, the language faculty consists of nothing but *Merge*, a binary structure-forming operation, which can produce complex syntactic objects by combining simpler syntactic objects, provided that the resulting structure meets certain constraints of readability by the semantic and phonological interfaces and that the processes involved are maximally computationally efficient. While the productions of human language appear full of quirks and irregularities, these should not be seen to reflect underlying complexity in the language faculty, but instead to result from the interaction of this system with all the others that are recruited in the production and interpretation of utterances. Because the distance between this faculty and behavior is so great, and the path from one to the other strewn with distorting influences, the hope is that surface complexity is consistent with deep simplicity. The difficulty with this view is, as Chomsky notes, "All the phenomena of language appear to refute it." (Chomsky and McGilvray 2012 p. 124). As illustrated by the cases above, the simple rules posited appear to be violated in a wide variety of different areas.

The various distinctions I have discussed, competence versus performance, grammaticality versus acceptability, and core versus periphery, play essential roles in justifying such simple descriptions. Each of these distinctions aims to draw a line between the genuine target of linguistic theory and various sorts of distorting influence. They provide space between theory and data: the theory aims to describe competence and grammaticality, while performance and acceptability judgements provide the data. Likewise descriptions of the core can be simplified in virtue of many of the complexities of linguistic behavior being viewed as peripheral.

As a result of this gap between data and theory, this approach has often been viewed suspiciously. Linguists, psychologists, and philosophers of quite different stripes have all argued that the practice of ignoring apparently pertinent data undermines the empirical credentials of generative linguistics. For example:

- "It is now evident to many linguists that the primary purpose of the [competence/performance] distinction has been to help the linguist exclude data which he finds inconvenient to handle." Labov (1971).

- "Since the actual linguistic conduct of any actual human language-user can be rendered compatible with any generative grammar by a suitable invocation of 'irrelevant performance factors', it becomes difficult to see how such a grammar could fail to be 'descriptively adequate' to a person's 'intrinsic competence'." Rosenberg (1988).
- "The idea of performance masking competence is also pretty much unfalsifiable. Retreats to this type of claim are common in declining scientific paradigms that lack a strong empirical base." Ibbotson and Tomasello (2016).
- "[Our] theory aspires to be accountable to all the facts, and not to be limited by a competence/performance or core/periphery distinction." Jackendoff and Audring (2019).

Each of these quotes points to the same sort of worry: if linguistics is to be a genuine empirical science, then it must be sensitive to observational data. But, it is argued, if the generativist is able to decide *post hoc* which data count and which can be ignored, this attitude results in a kind of confirmation bias.[21] The data correctly predicted by generative theory are viewed as genuinely relevant, while the problematic data can be dismissed as reflective of extra-linguistic influence (performance factors, peripheral influence, etc.) and therefore outside the scope of linguistic theory. The Galilean style, and the various distinctions it deploys, thus amounts to immunizing our linguistic theories from the observational data in a way that tarnishes their empirical credentials. If, it is argued, linguists are free to attribute any linguistic phenomena they observe to systems other than the system they are trying to describe, it will always be open to them to treat apparent counterexamples as irrelevant to their concerns. Successful predictions can be viewed as confirming their theories, while unsuccessful predictions can be viewed as irrelevant. The impossibility of falsification of the generativist proposals that this practice enables results in the theory generally being unscientific.[22] In the next few sections, I shall argue that this objection is mistaken, and that, if certain conditions are met, the Galilean style is a legitimate scientific strategy. Crucially, this will involve placing important constraints on the conditions under which data can be dismissed as outside of the scope of a theory.

5 A diagnosis

The problem that the Galilean style aims to solve is widespread in the sciences. It arises as a result of tension between several widely accepted claims:

Unification: One aim of science is to uncover the ways in which apparently disparate phenomena stem from the same sorts of processes.

Minimal Empiricism: Observations are the ultimate arbiter of the success of a scientific theory.

[21] Nickerson (1998) provides a general discussion of confirmation bias: the tendency to look for or focus on evidence favorable to one's antecedently held views.

[22] Note that 'falsifiability' here needn't be read in the strict sense according to which, due to the Quine-Duhem thesis, no theories are falsifiable. The real objection, and thus the one I take the explanatory economy approach to rebut, is that the various tools of the Galilean style create too great a distance between theory and data for proper empirical evaluation.

Complexity: Observable data are the products of the complex interaction of a number of interacting mechanisms.

Unification (as discussed by Kitcher 1981) is the view that one of the goals of a successful science is to maximize the explanatory coverage of a theory while minimizing its theoretical posits.[23] In particular, explanatory depth is achieved by showing how multiple phenomena result from the same underlying processes/principles. While I am skeptical that this could be the whole story about scientific explanation, it does seem that such an impulse characterizes much of the history of generative linguistics, among other disciplines, quite well.[24] A recurring theme in the work of Chomsky and those following him is that human language is far less diverse than it initially seems. That is, that the underlying similarities between all the world's languages are actually much more significant than the *prima facie* differences. Especially within the current Minimalist Program, generative grammar aims to show how much can be explained with very little.

Minimal Empiricism is the claim that ultimately, scientific knowledge must be grounded in observation. Of course, exactly what it means to be 'grounded in observation' is a hugely fraught issue, and there is no denying that the relation between confirmation and observation can be immensely complex, but the driving intuition is that scientific theorizing must be responsive to our observations. This claim has been central to all mainstream work in contemporary philosophy of science. Even theories of scientific theory confirmation that stress the importance of factors other than empirical adequacy (e.g. Kuhn 1962/2012 or Longino 1995) retain a central role for observational data. The standard picture affords these non-empirical virtues a role primarily in adjudicating between theories which are empirically equivalent.

The puzzle arises by combining these two meta-theoretic claims about scientific inquiry with Complexity. This is the claim that the observable data are typically the product of the interaction of numerous underlying systems. The problem is that complex systems typically display highly variable, context-sensitive behavior. In our case, despite allegedly having the same Universal Grammar, linguistic behavior varies significantly from population to population, and even person to person. This creates a dilemma for linguists: unified theories are unlikely to capture this variance, while empirically adequate theories are unlikely to be very unified.

The disagreement between generativists and the critics raised in the previous section stems from different responses to this tension. Generativists typically, in line with the Galilean style, seem to de-emphasize Minimal Empiricism, for the sake of Unification. By viewing only some of the data as relevant, the empirical difficulties (i.e. anomalies) with unified proposals are minimized. Opponents of the Galilean style, such as

[23] Strictly, Kitcher's account focused on minimizing the *argument patterns* of a theory, not merely its theoretical posits. However, as argument patterns are individuated partly by the non-logical terms used in their derivations, increasing the number of theoretical posits necessarily increases the size of the explanatory store (the set of argument patterns accepted as legitimate by the theory) and thus decreases unification. This is crucial for Kitcher's solution to certain problems raised with deductive accounts of explanation. For example, in order to show why explanations citing irrelevant properties do not count as genuinely explanatory he assumes that explaining the dissolution of a sample of salt in water with reference to its being hexed salt and with reference to its being salt use *different* explanatory patterns, and thus that an explanatory store containing the former will necessarily be less unified than one containing the former.

[24] See Freidin and Vergnaud (2001) for a nice discussion of the role of unification in generative theory.

usage-based or construction grammarians (e.g. Goldberg 2006 and Tomasello 2009), instead hold Minimal Empiricism firm, and thus reduce the push for unification.[25] Such theorists are typically led to view language as a highly particularized aspect of human culture, view linguistic diversity as deep-seated, and thus view the explanation of linguistic phenomena, and their acquisition, as much more of a piecemeal project.[26] The de-emphasizing of Minimal Empiricism by the generativists is often viewed as leading them towards a purely formal enterprise, rather than a genuinely empirical one.

6 The explanatory economy

In this section, I further explicate the notion of an *explanatory economy* and show how it can serve to justify the Galilean style. By recognizing our linguistic theories as embedded in such an economy, we can adopt the Galilean style without rejecting or de-emphasizing Minimal Empiricism. This approach thus dissolves the tension discussed in the previous section.

The thought, expressed in the quotes in Sect. 4, that treating a datum as outside the scope of one's theory necessarily diminishes a theory's empirical credentials stems, I believe, from an over-emphasis on *individualistic* understandings of theory evaluation, and an under-emphasis on *collectivist* understandings. An individualistic model of theory evaluation in the sciences treats only those observations that a theory makes predictions about or provides explanations for as relevant to confirmation. A collectivist model, however, also views the phenomena that a theory *does not* cover as relevant to confirmation. In particular, if the phenomena claimed to be outside of the scope of a theory are generally *within* the scope of another theory, compatible with the first, this counts in favor of the former theory. In this way, the confirmation of one theory is dependent *both* on the data it can cover *and* the data that other theories can cover.

No theory explains every observation. Even the best linguistic theory will tell us nothing about why, say, storing bananas together reduces ripening time. However, some data seem *prima facie* like a good linguistic theory would explain them. The distribution of wh-expressions in acceptable English sentences seems like a prime case here. However, when a phenomenon of this sort is *not* explained by linguistic theory, theorists have two options. Either they can modify the theory so as to explain the phenomenon in question, or they can claim that our original intuition was incorrect and that explanations of these phenomena are in fact outside of the scope of this theory.[27] Moves of this latter sort—of excluding some intuitively pertinent data as outside the scope of the theory—are paradigmatic examples of incurring explanatory debts.

[25] Perhaps the most extreme advocates of this position are those computational linguists heavily influenced by work in Artificial Intelligence, such as Steven Abney, Christopher Manning, Peter Norvig, and Fernando Pereira. See e.g. Manning (2003) for discussion.

[26] See Evans and Levinson (2009) for a clear statement of this position.

[27] Fodor (1981) provides an early argument for the view that which data linguistic theory ought explain is itself an *a posteriori* matter.

The status of observations that are claimed, contrary to our intuitions, to be outside the scope of the theory (such as echo-questions) is not the same as that of observations that intuitively fall outside the proprietary domain of a particular theory (such as the example involving bananas). In the latter cases, there is no presumption that linguistic theory should provide explanations. However, in the former, there is such a presumption, and therefore excluding it from the purview of one's theory, i.e. incurring an explanatory debt, requires motivation. Linguistic theorizing begins with some roughly identified collection of phenomena in need of an explanation. As theory develops, it becomes plausible to think that some of these data are better explained by other theories.[28] Given that we'd like an explanation of our facility with echo-questions, if (generative) grammatical theory is not going to provide such an explanation, some other theory, of something other than core grammatical competence, had better be able to. This is the sense in which a debt is incurred. Generative grammar's immunity from the apparent counterexample provided by echo-questions is bought at the cost of depending on some other approach to explain these phenomena.

Of course, the best reason to accept the view that some phenomenon need not be explained by a particular theory is that it has in fact been explained by a distinct theory. In such a case I will describe an explanatory debt as being *discharged* or *paid*. However, more frequently in actual science, debts can be evaluated based on how *likely* they are to be discharged. Whether some phenomenon is indeed profitably thought of as outside of the scope of a particular theory is itself a question that scientists can assess. In some cases, this may seem quite plausible while in others it may seem quite implausible. Cases can be made for the legitimacy of incurring some debt by suggesting an area of inquiry in which it can be explained, suggesting a style of explanation along these lines, and perhaps drawing analogies to other phenomena explained in this area, all before this explanation has been provided.[29]

This collectivist approach thus provides a way of thinking about counterexamples unlike that found in the main strand of theorizing about confirmation in philosophy of science. Rather than indicating that the theory (or the auxiliary assumptions) must be modified, at least some counterexamples point to the need for a division of labour among theories.

A proviso: I said earlier that for one theory to discharge the debts of another, these theories must be compatible. Exactly what 'compatibility' here amounts to is vague. Of course, logical consistency is required: it would do no good to defend a theory by showing that a logically incompatible theory is capable of explaining away the anomalies it raises. However, compatibility in a more general sense is desirable. Very generally, two psychological theories are compatible if they could both be true of the same cognitive system. In the cases I am focusing on, I believe an assumption of compatibility is warranted by the fact that the two theories target distinct systems

[28] It is also likely that data that initially seemed to be covered by another domain will be best explained by linguistic theory. Purported examples of linguistic effects in broader cognition, such as the role of language in the development of human Theory of Mind (see e.g. Astington and Jenkins (1999)), or in our ability to reason about mathematics or space (see e.g. Spelke (2003)) may be examples of this.

[29] Very likely, the notion of 'discharging' a debt will be a gradable one, with debts being more or less fully discharged by other theories. This is simply because the notion of explaining a phenomenon is similarly gradable.

(grammar and parsing systems in my first case study, and core and peripheral systems in my second). I am most concerned that this condition rule out debts generated by one theory of a particular target being discharged by distinct theories of that same target. For example, a grammatical theory which posits only binary branching structures (e.g. Kayne 1994) could not have any debts it incurred paid off by a rival theory of grammar which posits multiply branching trees (e.g. Culicover and Jackendoff 2005).

Beyond these 'intra-discipline' cases, it will typically be a difficult empirical question whether two theories are compatible in this sense, and it is unlikely that there will be any general answers to what compatibility requires. In some cases, there may be a clear line of inquiry that could answer this question. For example, if two psychological theories are of targets which interact with one another, we can ask whether the proposed outputs of one system are 'readable' by the other: theories of a parser and a grammar which output representations in completely unrelated formats are unlikely to be true of the same system.[30] But the further apart the systems are, the less we will be able to use this heuristic: ultimately a theory of grammar may need to have its debts paid by a biophysical theory, but as of now this provides very little in the way of guidance as to whether these theories are compatible. Relatedly, it is often an empirical question whether two theories should be viewed as competing or complementary accounts of the same phenomenon. A famous example of this involves the debate about whether connectionist architectures should be viewed as radical alternatives to symbolic systems (as e.g. Churchland (1996) has argued) or implementations of these traditional systems (à la Marcus (2001)). I take it that all of these debates are essentially empirical issues, not suitable to *a priori* determination. If even the compatibility of seemingly radically different architectures cannot be determined in advance, it seems unwise to propose strong criteria for when theories of different cognitive domains are compatible.

Overall, the constraint that debts must be discharged by compatible theories is analogous to the claim that successful explanations must have true premises: it is probably correct, but often will not be of much help. We simply have to wait and see how the empirical work turns out before we can determine whether this constraint has been met, except in the simplest of cases. Much more could, and should, be said about the conditions under which two theories or models are compatible, but I will leave this question here as this would take me too far afield.

There are at least two ways in which a theory can accrue a debt. Firstly, a debt is accrued when some observation that intuitively falls within the scope of a theory is excluded from the purview of this theory. Incurring such a debt is legitimate to the extent that another theory, compatible with the original, is able to explain this observation. This is the case I focus on in responding to the worries about the empirical status of generative linguistics. Secondly, a debt can be accrued by making some assumption the truth of which is to be verified by some other theory. Incurring these debts is legitimate to the extent that these other theories are able to provide this verification. In

[30] A further complication here is that there is a debate about the extent to which parsers utilize grammars in parsing, or whether they utilize independent rules which produce structures broadly in-line with the rules of the grammar, which may then be tested for grammaticality by the grammatical system. See for example Phillips (2004) and Stabler (1991) for arguments that parsers utilize the rules of the grammar, and Ferreira and Patson (2007) and Townsend and Bever (2001) for the opposite approach.

Sect. 9, I shall discuss such cases. In general, explanatory debts correspond to dependencies between theories. When the success of one theory relies on the explanations of another, the first is indebted to the second.

We can think of the confirmation of a theory as partially conditional on the set of explanatory debts it has accrued. Thus, the more unpaid debts a theory has, the less likely it is to be correct, but as these debts get discharged the plausibility of the theory increases. Debts will also vary in magnitude: the accrual of some kinds of debts will put a theory in a highly precarious position and ought thus motivate a serious effort to discharge them. On the other hand, some debts will seem much less pressing. Just as in the case of determining how much we ought increase our confidence in a theory given a particular prediction/explanation, determining what makes a debt large or small is highly impressionistic.[31] In the same way that failed predictions that seemed relatively unimportant can eventually overturn a theory (as in the difficulties encountered explaining Mercury's perihelion progression by Newtonian mechanics), debts that initially seem relatively insignificant can become highly important over time. This is just to say that judgments about the magnitude of a debt are, like all scientific judgments, fallible.

That said, there are a few clues as to how seriously particular scientists should view the debts their theories accrue:

Repeatability: When the same anomaly keeps arising, it is more in need of explanation. Given the complexity and context-sensitivity of the targets of linguistic, there is always the possibility that weird things might happen due to some unknown confluence of factors.[32] However, when the same weird thing happens in multiple different systems, or repeatedly over time in the same system, this indicates a systematic phenomenon. As linguistic explanation involves citing rules/generalizations, systematic phenomena are more plausibly within their purview than one-off events. Showing that this systematic, but linguistically unexpected, phenomenon can be explained by a distinct theory undercuts the possible need to modify our linguistic theories to account for it.

Robustness: The magnitude of a debt is increased even further when the same kind of anomaly arises in a wide variety of different circumstances.[33] Even more strongly than repeatability, the robustness of a phenomenon is an indication that it is systematic, and thus needs an explanation of the sort that linguistic theory can provide. In addition, robust phenomena restrict the kinds of explanations that can serve to discharge debts, as only systems that are uniform with respect to the wide range of circumstances in which these observations have been made can be used to explain these phenomena. For example, on the assumption that human linguistic envi-

[31] Even the most widely accepted, and most successfully formalized, account of theory confirmation, Bayesianism, is subject to certain fundamental difficulties, such as the problem of old evidence and the problem of the priors (See Earman (1992), especially chapters 5 and 6, for discussion). It seems that fully precise accounts of scientific justification and reasoning are liable to be perpetually stymied by the complexity of actual scientific practice. So, while certain important features of the explanatory economy are relatively underspecified, and would benefit from precisification, it seems likely that some degree of vagueness is irreducible.

[32] Wilson and Peters (1988) presents some amazing anomalies which seem to be of this kind.

[33] See Weisberg (2006).

ronments are less similar across agents than human biology is, species-universal properties are better explained with reference to innate traits than acquisition. Linguistic universals, if there are any, are thus suitable targets of a fairly narrow range of theories, and so if they are not explained by a linguistic theory, the debt incurred will be quite substantial.

Entrenchment: Some of the properties of target systems are relevant to the explanation of a wide range of observable phenomena, whereas others are relevant only to a few.[34] Anomalies out of step with the predictions of claims about these central systems are more critical than those that suggest mistaken views about less entrenched phenomena. Entrenched properties reverberate throughout the system, and so mistakes in this area lead to theories that are not merely inadequate but wrong-headed. Theorists must therefore be extra sensitive to any anomalous predictions that stem from descriptions of entrenched systems/aspects. This makes debt collection in this area particularly pressing. In linguistic theory, 'architectural' properties of the grammar, such as the constraint that trees be binary branching, are relevant to all linguistic phenomena, and so apparent difficulties with these properties will require more immediate solution than more particularized issues relevant only to certain constructions.

Centrality: Theorists often view particular phenomena as particularly indicative of the system they aim to understand. Failing to account for these phenomena creates a more pressing problem for these theorists than failures to account for more peripheral phenomena.[35]

This list is of course very preliminary in both detail and range, and each factor is as vague as the notion of debt magnitude it is intended to elucidate. I am sure there are more factors than just these four that go into determining how serious a debt is, and much more must be said about each of these four features. I mention them just to give a flavor for the kinds of factors I have in mind and hopefully to spur future research.

In its non-extended sense, an economy consists of a system of producers and consumers. Consumers have demands for certain kinds of resources, which can be supplied by the producers. The complex structure of such a system, and the constituent members thereof, determine the distribution of these resources. Idealizing: we can treat scientific theories/programs/models as producers of explanations.[36] My claim is that traditional theories of confirmation have acted as if the success of a particular program can be evaluated by viewing the product of this program alone, by comparing the explanations produced with the data to be explained. However, this ignores the interdependence of these producers. In the cases I am interested in, one producer 'out-

[34] See Wimsatt (2007).

[35] Centrality is a particularly hazy notion, even more so than the others. However, I am confident that anyone who studies scientific practice will recognize it. Some phenomena just *feel* like a theory in this area must account for them, whereas others seem more marginal. Island constraints and binding principles had better be accounted for if a linguistic theory is going to be worth its salt; heavy-NP shift and appositive relative clauses, perhaps less so.

[36] I have very little to say about the consumers in such a system, but these will include all those systems that make use of scientific explanations, be they governments producing scientifically-informed policy, manufacturers of technology, or members of the populace seeking knowledge. Kitcher (2011) provides an extended discussion of the demands of such consumers.

sources' the production of a particular explanation to another. The assessment of the overall success of the former program will thus depend on the success of the latter.[37] Likewise, the success of the latter program will be increased in virtue of its paying off the debt of the former. Not only does the production of an explanation of this anomalous datum increase the predictive/explanatory coverage of the theory, but it also shows how well this theory meshes with the others.

We can now revisit the diagnosis of the dispute between those who adopt the Galilean style and those who reject it. I said above that the Galilean style involved de-emphasizing Minimal Empiricism in favor of Unification. This can now be stated a little more precisely, by distinguishing between *local* and *global* versions of Minimal Empiricism. The local version says that, for some specified class of data, an empirically successful theory must be consistent with these data. For example, an empirically successful grammatical theory must, among other things, successfully predict the distributional facts about wh-expressions. From this perspective, generative grammar is failing, as echo-questions provide counterexamples to its proposals. I believe the skepticism about the empirical status of generative grammar is motivated by this local approach. However, a global construal of Minimal Empiricism says instead that the data must be accounted for, but doesn't commit particular theories to accounting for any particular data. From this perspective, the phenomenon of echo-questions (and unacceptable center-embeddings) must be explained by some theory, but it need not be a theory of (core) grammar. The explanatory economy approach shows how the Galilean Style is consistent with global, but not local, Minimal Empiricism. By excluding data from the scope of generative grammar, the local version of this constraint is violated. But by ensuring that some other theory accounts for these data, the global version is maintained. The hope then is that a multitude of mechanisms/systems can be described in relatively simple and general ways, enabling unifying explanations of particular phenomena, and that *collectively* these can account for much of the data, although prior to actually producing such explanations it may be difficult to discern which phenomena are to be explained by which theory.[38]

It is crucial here to note the differing roles that explanatory and predictive power play in evaluating theories.[39] The *aim* of a theory within cognitive science is to explain the workings of the mind and its role in behavior. Prediction, or consistency with the data more generally, provides guidance in determining which explanations are

[37] Terminological note: I realize that the 'debt' metaphor is slightly misleading, as the cases on which I focus are all ones in which the debt incurred by one theory is paid off by another. Were it that the real economy worked in this way. However, I believe this is a product of my focus, rather than the nature of the phenomenon itself. There is no reason why a debt cannot by paid off by the theory that incurs the debt. This is just the normal case of a theory making a bad prediction and then going on to resolve this issue. I focus on the cases of debts being paid off by other theories because they are less well understood.

[38] Johnson's (2009) discussion of semantics suggests a similar picture, according to which semantic theory aims not to account for all (or perhaps even any) observed data, but instead to provide one source of variance, which may eventually be supplemented with other sources which can collectively account for all (or at least most) of the data.

[39] Of course, theories can be evaluated on many more dimensions than these, including how fruitful they are in provoking novel and interesting ideas, how much they contribute to the production of societally beneficial technology, etc. My aim here is to assess the purely epistemic arguments for or against generative linguistics, and so will restrict my attention to explanation and prediction for the purposes of this paper.

correct. For example, a linguistic theory aims, among other things, to explain why speakers react differently to sentences like (4) and (5) above. It does so by positing grammatical rules which make the former grammatical and the latter ungrammatical. However, this is not yet sufficient to predict behavior. Remember that it is acceptability, not grammaticality, which is observed in linguistic judgments. On the *assumption* that (un-)grammaticality leads to (un-)acceptability, these theories can then issue a prediction. It is these predictions against which the theory is tested. However, when such a prediction fails, it can be argued that this assumption is not met, and so the theory in fact makes no prediction at all. Explanatory debts function to differentiate apparent anomalies, in which no genuine prediction is made, from genuine counter-examples, in which a prediction fails.[40]

One potentially counter-intuitive feature of the view here developed is that the unification of a theory is increased by allocating observations to the explanatory purview of another theory. This seems like it would *decrease* the unification of the theory, in that it decreases the theory's explanatory scope. And indeed, on Kitcher's account, unification is increased by increasing the set of observations a theory can account for. However, this is not the only thing that matters for unification: the unification of a theory is likewise increased by reducing the number of explanatory patterns needed for these explanations. The explanatory economy is a response to the fact that these two criteria trade off against one another: in many cases, greater explanatory coverage is made available only by increasing the set of explanatory patterns. In the cases described above, one could modify the grammar so as to explain the recalcitrant phenomena (e.g. by adding rules constraining center-embeddings), but to do so, one would have to add to the explanatory store and thereby reduce the unifying power of the theory. The lesson of the explanatory economy is that it is often better to develop multiple theories, each of which requires a very few explanatory patterns, capable of collectively explaining the observations, than to propose a single theory with a very large explanatory store. This is again a reflection of the distinction between individualist and collectivist accounts of theory confirmation. One primary motivation for this collectivist approach is the power it provides in limiting redundancy. When solely focusing on a particular theory, it may be unclear whether we are better off complicating the theory to increase its explanatory scope or keeping it simple and excluding some observations from its purview. However, when we expand our focus so as to include the successes of other theories, decisive answers to these questions can be given. If the addition of an explanatory pattern to our theory enables us to explain certain observations, but these observations are already explained by a distinct, compatible, theory, then we have reduced the unification of our theory for no real gain. In the case discussed in the next section, modifying our grammatical rules so as to explain the unacceptability of multiply center-embedded sentences may superficially seem to increase the explanatory power of our grammatical theory, but given that a parsing

[40] This is related to the claim in Cummins (2000), that the central task of cognitive science is not to capture the behavior of an organism (by detailing various 'effects', such as the McGurk effect), but to provide broadly mechanistic explanations of why these behavioral patterns are observed. Note also that the separation between what a theory explains and what it predicts, due to the number of extra-theoretical assumptions needed for the latter, is familiar in other special sciences, especially evolutionary theory. See e.g. Scriven (1959) and Sober (1984).

theory is able to account for this phenomenon anyway, collectively our theorizing is no better off. This explanatory redundancy is avoided by keeping our theories' explanatory scopes restricted and disjoint, enabling sparser explanatory stores. In this way, I view Kitcher's account as closely related to that of Chomsky (2002), who argues that explanatory power is purchased by showing how multiple distinct phenomena can be accounted for by a single mechanism.

One related response that may suggest itself is: why should we continue dividing up the observations between theories, rather than simply combining the theories into one more general theory capable of accounting for all the observations? That is, rather than apportioning observations O and O' to theories A and B respectively, why not combine A and B into a larger theory C, capable of explaining both O and O'? This would enable us to retain our individualistic account of confirmation, as we could just evaluate C, rather than evaluating A partially on the basis of the successes of B, and vice versa. Likewise, this would avoid the worry about explanatory redundancy, as any observations apparently creating explanatory debts for A would be explained not by adding some explanatory pattern to A capable of replicating the successes of B but instead of simply incorporating the explanatory patterns in B responsible for these successes into C. This worry may seem particularly pressing given the constraint that theories A and B must, if one is to discharge the debts of the other, be consistent: if they are consistent, why not combine them?

There are, I think, several problems with this proposal. Firstly, since at least Kitcher (1984), the assumption that scientific theories can be identified with something like a set of sentences or propositions, as opposed to something much more amorphous, such as a "subject matter and…methods of investigation" (p. 340), has received significant scrutiny. But if scientific theories are individuated in this way, it is unclear what it would even mean to combine or unify them.[41] More generally, when scientists talk of producing a 'unified theory of mind', they do not mean merely the conjunction/combination of a variety of distinct psychological theories. There are psychologists pursuing such a project (e.g. Anderson's (2009) ACT-R architecture), but what they are aiming at will look quite different than the proposed unification of the theories in question, for a variety of reasons. Firstly, the examples I focus on of the explanatory economy involve pairwise dependence: a debt created for one theory is paid off by another. However, it is often unlikely that any real psychological theory would target these, and only these, two systems. For example, while debts accruing to grammatical theory can be discharged by a parsing theory, it is unlikely that any psychological theory will include accounts of grammatical and parsing systems and nothing else. Proposed unified cognitive architectures will include also perceptual-motor systems, memory systems, belief/desire cognition, etc., as well as accounts of the interactions between each such component. That is, the practice of the cognitive sciences suggests that theorizing is done with respect to individual systems or whole cognitive architectures, whereas the role of the explanatory economy is most useful in the intermediate level, focusing on small collections of theories of these individual systems, rather than the

[41] Note that this worry involves skepticism of both the syntactic and semantic views of theories. If theories are sets of sentences, as the former advocates, or sets of interpretive models, as the latter advocates, unification can be understood fairly straightforwardly as the conjunction or union of these objects.

entire mind. One could *call* some combination of a theory of grammar and of parsing a single theory, but this would be merely a verbal maneuver with little resemblance to the theories of cognitive science.[42] One could well think that the ultimate *goal* of (cognitive) science is to produce such a unified theory, including theories of all the components of mind and their interactions, but as no remotely adequate theory of this kind is in the offing, we seem now forced to be content with multiple distinct but interdependent theories, related by the explanatory economy.

The explanatory economy points to an important distinction between kinds of confirmation. A theory is directly confirmed by the observations it predicts/explains, and indirectly confirmed by the discharging of its debts. This latter kind of confirmation works in several ways. Firstly, when what seemed to contradict the theory is shown to be explicable by some other theory, an empirical problem for the theory is resolved: the theory's *empirical adequacy* is demonstrated. Secondly, when the exceptions to, or presuppositions of, one theory are explained by another, the *external consistency* of the theory is demonstrated: it has been shown that the theory not only explains its proprietary observations but these complement those of other disciplines. Thirdly, it demonstrates the theory's *fruitfulness*: the occurrence of anomalies to one theory has indicated fruitful research questions to be asked in other domains (as when center-embeddings, posing an apparent problem for grammatical theory, stimulated work for a theory of parsing). The role of explanatory debts in promoting these three virtues (lifted from Kuhn 2010) suggests that this is genuinely confirmatory, and is so in a way distinct from standard predictive/explanatory success. Simply combining the indebted and the debt-paying theory into one super-theory would blur this important distinction.

While I take the explanatory economy to provide a defense of the generativist methodology, it will require some re-evaluations of their strategies. Generativists tend to be fairly blasé about ignoring counterexamples. Cooper's (1983/2013) claim (p. 149), in a classic textbook, that "Accounting for echo-questions with the grammatical rules for questions would, I think, make it very difficult to give a general account of questions." is, I believe, fairly representative of the view of many mainstream generative linguists. It may be true that modifying one's syntactic theory so as to account for echo-questions would indeed be unprofitable. A *grammar* which specified, as well as the rules needed to explain mandatory wh-raising, all the ways in which certain discourses license violations of these constraints would likely be too complex and specific to be illuminating. But, this does not mean that such data can be totally ignored. The distribution of wh-expressions, unlike the ripening rate of bananas, intuitively fits squarely within the purview of syntactic theory. If these apparent deviations from the syntax of the language faculty cannot be explained by some other theory/discipline, then the debt remains unpaid and generative theories must face the possibility that this is a genuine counterexample. Generativists must therefore be attentive to those theories that are applicable to other aspects of the complex system responsible for linguistic production as they are explanatorily indebted to such approaches.

[42] This worry is especially forceful when the debts are discharged by quite different theories, as when a debt incurred by grammatical theory is paid by biological/developmental theory (see Sect. 9). In such a case, it strikes me as radically implausible to call the combination of two theories itself a theory.

I thus take the explanatory economy picture to be partially descriptive and partially prescriptive. Those in the generativist tradition do sometimes follow this methodology. That is, they exclude some phenomena from the purview of their theory (as reflecting the influence of performance systems or the periphery) and then go on to give reasons why these phenomena are plausibly explained by theories of systems outside of the scope of linguistic theory. The discussion of center-embeddings in the next section provides a particularly clear example of this. However, they sometimes shirk this duty and wield the competence/performance and core/periphery distinctions without giving due care to whether the debts they thereby incur are likely to be discharged. In such cases they are not following the strictures of the explanatory economy and so the charge that they are simply immunizing their theory from apparently inconvenient data may be warranted. If I am right, utilizing such distinctions is legitimate only when there is at least a prospect of such alternative explanation, and it behooves generativists to establish that this is the case.

According to this picture, then, we can find a path between the hyper-sensitivity to data characteristic of Popper or the Logical Empiricists, which precludes the proper development of theories, and the hyper-insensitivity suggested by simple characterizations of the Galilean style. On my account, all of the data are relevant, and must be accounted for.[43] This global Minimal Empiricism establishes enough of a connection to the data to rebut worries about being unscientific. But which data are accounted for by which theory is an open question. This supplies the space between data and theory necessary to allow deep and unifying explanatory theories to develop. Such a picture is necessary when attempting to theorize about complex systems with multiple interacting subcomponents responsible for the observable phenomena.

7 The explanatory economy at work

Let us now see how we can apply this methodology to the case studies described earlier. It was claimed that multiple center-embeddings were grammatical despite being unacceptable. The worry was that this claim amounted to little more than stipulating a gap between observation and theory, insulating our theory from the observations in a way inconsistent with the Minimal Empiricist constraint. The explanatory economy strategy claims instead that excluding data from the purview of our theory amounts to claiming that they are in the purview of some other theory, to which we are now indebted. If some other theory, of a different target, is capable of explaining these data, then this exclusion of this observation from the domain of observations for which grammatical theory is responsible is legitimized. In this case, just such an explanation is available.

Miller and Chomsky (1963) provide a computational/mathematical explanation for the difficulty generated by parsing multiple center-embeddings. To see how this explanation works, imagine a machine which aims to determine, for a given sentence, which

[43] Although, within certain bounds, some anomalous data (genuine noise) will remain unexplained. Crain and Thornton (2000), for example, suggest the heuristic that up to 10% of the data can be discounted as 'noise', while anything greater than that which conflicts with a theory's predictions must be accounted for by identifying the source of error (p. 45).

arguments (e.g. NPs) are assigned to which verbs. Such a machine, when presented with a sentence, moves from left-to-right, identifying each word and its part of speech. When an argument is identified, a 'file' is opened, only to be closed when the relevant verb has been identified, and the former assigned to the latter. In sentences like (1) and (2), when the first NP has been identified, it cannot be assigned to the main verb until it reaches the end of the sentence. Along the way, it identifies two more arguments, and so must open three files simultaneously. This creates too great a burden on the memory of the device, and so the sentence cannot be parsed correctly. In shorter (e.g. "The mouse the cat chased squeaked.") or right-embedded sentences (e.g. "The cat chased the mouse which squeaked.") the arguments can be assigned to their verbs without placing this burden on memory, and so they are interpretable. The constraints on this model (finite memory, left-to-right parsing, etc.) are near-certainly properties of the human parser.[44] If the human parsing system in fact has these properties, this explains why humans find interpreting center-embedded sentences difficult, and it does so without claiming that center-embeddings are *ungrammatical*. Of course, it is an empirical claim that human parsers work in these kinds of ways. If it were discovered that they do not, the status of the debt created for the grammar by center-embeddings may need to be re-evaluated.

This example shows the explanatory economy at work. A theory was proposed, but it seemed to face counterexamples. Modifying the theory so as to avoid these counterexamples seemed theoretically undesirable, and so the data were excluded from the purview of the theory. From one perspective, this may seem to be a bad result: reducing the scope of a theory so as to avoid counterexamples is supposed to be exactly what degenerating research programmes do.[45] However, this perspective stems from an overly individualistic view of confirmation. From a collectivist perspective, this move *increased* the explanatory power of cognitive science overall, as it led to the creation of a model, of a distinct aspect of the overall cognitive system, which explained the phenomenon in question and thereby showed *why* a grammatical model should not be expected to predict this sort of behavior.

The competence/performance distinction (and by extension the grammaticality/acceptability distinction) has often been maligned by those outside of the generative tradition—as seen in the quotes in Sect. 4—as making the theory unfalsifiable. If, it is argued, any apparent counterexamples can be simply dismissed as performance effects, outside of the scope of the theory, then the theory's credentials as genuinely empirical are reduced. The explanatory economy account shows why this worry is mistaken. By classifying a phenomenon as a mere performance datum, generativists incur a debt, but this debt can be discharged by looking at nearby disciplines, as in the explanation just discussed. Rejecting a counterexample as a performance effect, and thus incurring an explanatory debt, insulates theory from data *only if* this counterexample is also viewed as outside the scope of theories of the performance systems. If this datum is plausibly explicable by some other theory, and the proposed explana-

[44] A parser is a system that takes as input a physical token, say a sound, and outputs a structural representation. So for a sentence to be acceptable, the parser must be able to assign a structure to it. A grammar, on the other hand, determines which structural representations are grammatical and which are not.

[45] See Lakatos (1976).

tions of this datum by these other theories are themselves empirically testable, then the unfalsifiability worry misses the mark.

It is interesting to note that this early recognition that these debts should be paid seems to be in conflict with some of the recent statements of the Galilean style. For example, in his opening remarks to Piattelli-Palmarini et al. (2009), Chomsky describes this style: "You just see that some ideas simply look right, and then you sort of put aside the data that refute them." (p. 36). My argument is that the early methodological stance towards anomalies is entirely correct, and that simply 'putting aside' these inconvenient data, rather than investigating ways that alternative approaches may handle them, prevents generative approaches from maximizing their confirmation by appropriately assigning, and ultimately discharging, explanatory debts.

While generativists generally accept something like the explanatory economy picture when it comes to the competence/performance distinction, this attitude is less frequently seen in discussions of the core and the periphery. However, a justification for this categorization is likewise needed, lest it reduce to an *ad hoc* immunization of theory from data. Here again the notion of explanatory debts is useful. In treating echo-questions, for example, as outside of the scope of generative theory (of the linguistic core), these theorists incur a debt. If no other theory or discipline is able to account for our ability to use echo-questions as we do, then they may be re-considered as counterexamples to such proposals. But they may be at least temporarily discounted without either falsifying or undermining the empirical credentials of generative theory in the hopes of some future analysis. In this case, the situation isn't as good for the generativist as in the example of center-embedding. This debt has not yet, to my knowledge, been paid. But there are reasons to think that the incurring of this debt is reasonable, in that it is likely to be payable.

There are a variety of ways such debt-collection could go. One promising source for explanations of apparent counterexamples of this sort are usage-based or construction-grammar approaches.[46] For our purposes, what is interesting about these approaches is that they are 'periphery-first'. That is, they take the aspects of language deemed peripheral by generative grammar as the basic building blocks of language. The paradigmatic peripheral properties of a language are found in the lexicon. While grammatical principles/parameters may be universal and innate, which words are used is clearly neither. Chomsky (1995) goes so far as to *define* the lexicon as 'a list of exceptions' (p. 235). Construction grammarians, unlike generativists, view the acquisition of syntax and of lexical items as exemplifying the same mechanisms. Learning a language involves recognizing meaning-form correspondences of various types: first words and then larger constructions are acquired via a process of generalization and abstraction. For example, through exposure to sentences like 'I run' and 'you run', the child can learn that uttering 'NP run' is a way of conveying that some object runs. Combining this with their knowledge (acquired analogously) that 'NP walks' is a way of expressing that some object walks, they generalize to the knowledge that 'NP Vs' is a way of expressing that some object performs some action, and so on for more complex examples. These complex constructions are then stored in the same way that more

[46] I will use these terms interchangeably in what follows. For paradigmatic examples, see Goldberg (2006) and Tomasello (2009).

traditional simple lexical items are. In this way, linguistic competence in general can be viewed as the acquisition of more-or-less concrete form-meaning pairs, the more abstract of which provide the basis for the construction of novel expressions. From this perspective, the fact that the periphery is not *merely* a list of exceptions, but in fact includes significant amounts of its own structure, is taken as central.

Perhaps the most effective explanations in such a tradition involve precisely the cases in which the apparent violation of linguistic rules does not lead to unacceptability. For example, it appears that learning the adicity, or argument structure, of various lexical items is one of the central tasks of acquiring a language. Such knowledge explains why speakers treat sentences with the wrong number of arguments (e.g. *"Mikhail sneezed Bill") as unacceptable. However, in certain cases, these rules can be violated. Consider "Mikhail sneezed his tooth across the room" (see Goldberg (2006)) or "Mikhail sneezed his way through the meeting" (see Jackendoff (1992, 2010)). In such cases, it is argued that prior exposure to expressions with these structures ("NP V NP PP" and "NP_1 V NP_1's way through NP_2" respectively) gives learners the knowledge that such constructions are legitimate, *even when* the verb that features within them cannot normally be used with these additional arguments. In this way, the acquisition of information about complex linguistic expressions (constructions) enables speakers to over-ride other rules/constraints that form part of their linguistic competence.

Defenders of these approaches are typically ambitious in viewing them as sufficient to account for our linguistic capacities. They typically run a slippery-slope argument aimed to show that the abstract and innate grammar of the generativists is not needed, as the mechanisms by which the periphery is acquired are sufficient to account for the grammatical knowledge traditionally located in the core.[47] Once the need to attribute significant amounts of structure to the periphery is admitted, it is argued that *all* linguistic structure can be accounted for in the same way. However, I believe the methods they propose should be thought of as only part of the story about language acquisition. In particular, these processes of abstraction and generalization can be taken to be the mechanism by which peripheral phenomena and exceptions to core rules can be learned, thus providing a way of discharging the explanatory debts of the generativist. That is, I believe we should view the periphery and the core as both containing structure, but that these respective structures are acquired and utilized in quite different ways. Innate constraints on acquirable grammars—as described by the generativists—are found in the core, while linguistic patterns abstracted from the environment—as described by construction grammarians—are found in the periphery.

From the earliest days, generative linguistics was concerned with the puzzle of how a child can acquire a language given only the evidence made available by parents and peers in the child's environment. This problem is made acute by the poverty of the stimulus. The evidence available to the learner seems to characteristically underdetermine the acquired grammar. For example, adult speakers have clear intuitions about the grammaticality of expressions which are almost never found in the learning

[47] See e.g. Tomasello (2009) and Jackendoff (2018).

environment, such as parasitic gaps.[48] To account for this fact, the grammatical principles responsible for such phenomena are posited to be innately given. If children are born knowing the facts about grammar which determine when gaps are and are not licensed, then there is no need for them to learn this from the available evidence, and thus no worry about the sparsity of such evidence.[49] This poses a deep problem for construction/usage-based approaches, which view all language acquisition as extrapolation from observed environmental patterns. If the apparent constraints on the acceptability of certain expressions are typically absent from the linguistic environment of the learner, it seems that the acquisition mechanisms of these approaches are insufficient. I do not believe these approaches have adequately countered such arguments, and thus an adequate theory of language acquisition must include at least some reference to the innate core.

This suggests a relatively neat division of labour between generativists and other kinds of syntactic theorists. Classical generativist arguments involving poverty of the stimulus and unification (of disparate constructions and languages) can delimit a class of phenomena (the core) to which generativist theories are centrally responsive. If a phenomenon is present in developed competence, despite being absent from the learning environment, then it must be at least partly a product of the learner's innate endowment. This makes alternative (non-generativist) explanation implausible, and so such phenomena are not suitable for the invocation of an explanatory debt on behalf of generative grammar. However, to the extent that the acquisition of linguistic phenomena seems to be explicable via extrapolation from the learner's environmental linguistic data, they may be suitable targets of explanatory debts. Usage-based theorists have complex and subtle models of acquisition via abstraction and generalization, and so a divide-and-conquer approach seems viable. These alternative approaches can thus be seen as necessary benefactors of the generativist project, underwriting the latter's debts and showing how the Galilean style, with its focus on general, unified explanations, is consistent with (global) Minimal Empiricism. The proposals of the generativist can be viewed as describing the universal features of the language faculty, while empiricist models provide explanations for the quirks and irregularities that seem to pose empirical problems for these general proposals. This division, between the universal and unlearned core, and the variable and environmentally dependent periphery, provides a further guideline, specific to linguistic theory, for determining which explanatory debts are legitimately incurred, in addition to the very general heuristics discussed in Sect. 6.

Echo-questions seem particularly amenable to a usage-based theory of acquisition. These theories focus on general, i.e. not language-specific, features of human cognition in their explanations of acquisition phenomena. In particular, they highlight humans'

[48] Parasitic gap constructions are sentences in which an expression pronounced at one location is interpreted at multiple other locations at which it is not pronounced and for which the (phonologically null) occurrence at one of these other locations is licensed only when the expression also occurs unpronounced in the other. For example "Which books did you read [which books] and tell your friends about [which books]?" versus *"Which books did you read [which books] and tell your friends about movies?". Such sentences are almost never attested in adult corpora, let alone child-directed corpora. But nonetheless, competent speakers have fairly robust intuitions about the subtle and surprising patterns of acceptability they exemplify. See Engdahl (1983) for a classical discussion of these phenomena.

[49] For an excellent survey of such arguments, see Crain and Pietroski (2001).

unique capacities for socializing and social reasoning, such as their unusually high ability and desire to enter into episodes of joint attention and their extreme capacity to read the intentions of others.[50] This makes for a 'pragmatics-first' approach to language acquisition: expressions are acquired based on mechanisms for grasping what was intended by their usage. This creates puzzles about the acquisition of structures which seem removed from pragmatic motivations, such as the constraints on parasitic gaps mentioned above, where there seem to be perfectly clear messages that someone could intend to convey but these readings are unavailable. However, this approach is well-suited for the acquisition of constructions which serve a clear pragmatic role. Echo-questions seem to fit this bill nicely. These questions are acceptable only in certain specific discourse situations. In particular, they are acceptable only in response to an (immediately) previous assertion, and they serve only to question some aspect of this assertion.[51] For example, take sentence (5) above ("Xian will buy what?"). This sentence is acceptable only in response to an assertion like sentence (3) ("Xian will buy a car.") and is used to convey that either the asker did not hear what it was that Xian was going to buy, or that what Xian was going to buy was in some way inappropriate. That such expressions are used always in the same discourse role, and with broadly the same speaker intentions makes them particularly suitable for a theory of acquisition based on intention-reading. To my knowledge, there has not been a usage/construction-based account of echo-questions along these lines, but it seems particularly plausible.[52]

Crucially, the claim that echo-questions are a peripheral phenomenon, to be explained by something like the usage-based theory of language acquisition, is empirically testable. In this way, the generativist who incurs a debt here is not simply ending inquiry by rejecting an apparent counterexample, but is instead proposing further avenues for research. For example, if echo-questions are indeed peripheral phenomena, learned exceptions to the core rules of wh-movement, then they must be learnable from the environment. Usage-based theories are empiricist in that they view acquisition as a process of extraction of information from environmental patterns. This means that these models can only explain the acquisition of a certain subset of linguistic phenomena: those for which there is a suitable inductive base in the learner's linguistic environment. So, to see if this debt is dischargeable, we could look at something like the CHILDES corpus of child-directed speech, and see how frequently we find echo-

[50] See e.g. Tomasello et al. (2005).

[51] I am oversimplifying slightly here. The actual range of discourse situations which license such questions is complex and seems to include, as well as situations in which the queried claim has been uttered, situations in which the queried claim has been made salient through other means. For example, A: "My mother and brother graduated from Harvard." B: "And your father graduated from WHERE?". This complexity should not matter for my purposes, and I shall ignore it for the remainder of the paper.

[52] Keep in mind that this is just one plausible story. There are others, such as Karmiloff-Smith (1995)'s theory of representational redescription, according to which certain structures begin as informationally inaccessible subroutines of a system, but are 'redescribed' so as to be available as representational objects of computational systems which can then manipulate them. Theorists in the HPSG tradition, discussed in fn. 18, also have provided explanations for such phenomena. However, these are less amenable to collaboration with generativist models, as they provide an entirely novel syntactic framework. Yet another alternative is that echo-questions are an extension of certain kinds of primitive mimicry capacities. It is, of course, an empirical question which such approach ultimately provides the best account for this phenomenon. The point is just that such approaches are available and may potentially discharge the generativists' debts.

questions. If these are common, and occur in a variety of forms, it is plausible that children learn how to violate innate rules of wh-movement by hearing others do so. However, if we find that echo-questions are infrequent in child-directed speech, the claim that they are *learned* exceptions is made highly unlikely. That is, if there is a poverty of the stimulus for echo-questions, then they must be located in the innate core, and generativists may indeed have to revise their core grammatical theories to handle them. This is all, of course, quite sketchy, but hopefully it shows how the explanatory economy can further inquiry by showing how to determine which observations are to be handled by which theory.

Some such story must, for the core/periphery distinction to be consistent with Minimal Empiricism, be in place. When behavioral outputs are *prima facie* inconsistent with the rules that purport to govern the workings of the language faculty, there must, in order for the generativist theory to remain empirically viable, be some other system responsible for these deviations. My claim is that the empirical status of generative linguistics depends on an understanding, at least in broad outline, of such mechanisms and systems. This intermediate position thus finds space between those, like the usage-based theorists, who claim that the core/periphery distinction is unreal and appeals to it are unscientific, and generativists in certain moods who deny the importance of these alternative approaches to languages. These approaches need not be seen as rivals, but instead as targeting different aspects of the complex system responsible for human language.

This novel view of linguistic theory, as dependent on the combination of generativist theories of the innately-specified core and usage-based theories of the learned periphery, has, I believe, been neglected due to too much emphasis on the individualistic model of confirmation. When these theories are put forward, they are typically viewed as competitors: one must accurately describe the human capacity for language to the exclusion of the other. This leads to argumentative stalemates, as poverty of the stimulus arguments seem insurmountable for the usage-based approach, while the large amount of data outside of the scope of the relatively sparse generativist theories seems to provide empirical refutation of these theories. However, the collectivist, explanatory economy approach can make perfect sense of this. Both of these theories are important, and they have complementary domains of coverage. Each is needed to fill in the holes left by the other.

Hopefully this discussion has shown how fruitful the explanatory economy approach can be. Superficially, the competence/performance and core/periphery distinctions seem to proliferate the ways in which apparently relevant data are dismissible, thus making the theories utilizing them seem immune to counterexample. However, the explanatory economy approach systematizes these distinctions, showing how they involve the creation of explanatory debts. As long as the payment of these debts is taken seriously, the connection between theory and data, i.e. Minimal Empiricism, can be maintained.

8 Horizontal and vertical dependencies

The kind of explanatory pluralism for which I am advocating here is a kind of 'horizontal' analogue of the widely accepted division of labor in cognitive science between different levels of explanation. The most famous example of this comes from Marr (1982). Marr argued that psychology makes progress by explaining behavior at three different levels: the computational, algorithmic, and implementational levels. At the highest, computational, level it is specified what the psychological system is supposed to do. In particular, specifying this level involves determining the function-in-extension of the system—its mapping of system inputs to system outputs—and showing why, with reference to environmental regularities, computing such a function benefits the system (e.g. by showing that, under certain circumstances, computing a particular mapping from retinal stimulation to representations of edges tracks the relation between projected environmental light intensities and objectual edges).[53] At the intermediate, algorithmic, level this function-in-extension is specified as a function-in-intension, which makes explicit the stepwise, algorithmic, process by which the function-in-extension is computed via a series of operations applied to a specified class of representations. Finally, at the lowest, implementational level it is specified how this algorithm is instantiated in a physical (e.g. neurobiological) system.[54]

This approach to cognitive science involves a distribution of observations to be explained and questions to be asked, just as do the examples of the explanatory economy I have described. For example, in the linguistic case, a generative grammar provides a computational level description of human linguistic competence.[55] This is suitable for explaining which sentences are or are not part of a given language. However, psycholinguistic data, such as measurements of the time it takes for a subject to determine whether a sentence is well-formed or not, do not typically bear directly on the questions at this level. Instead, such observations can be explained at lower levels. Likewise, neurolinguistic data, such as the observation that a certain event-related potential is correlated with certain linguistic properties most plausibly bears on questions about the implementational level.

The examples I have focused on have all involved the collaboration of psychological theories at the same, computational, level: theories of linguistic competence having their debts paid by theories of memory or parsing likewise stated at the computational

[53] As this makes clear, the computational level combines two different kinds of question to be answered: *what* does the system do, and *why* does it do this. See e.g. Shagrir (2010) for a nice discussion of this point. I will largely be focusing on the what-questions, aimed at specifying the task that the system performs.

[54] Newell (1982)'s *Knowledge Level* and Anderson (2013)'s *Rational Level* are akin to Marr's computational level, in being the locus for high-level descriptions of the task being performed, with more detailed accounts of precisely how these tasks are implemented occurring at lower levels. I will focus on Marr's approach, as these alternatives are best suited for accounts of the behavior of whole organisms, whereas Marr's account is developed precisely to account for the behavior of subsystems. But what I say about Marr's account should be applicable *mutatis mutandis* to these alternative approaches.

[55] Actually, fitting generativist theory into a Marrian framework is a little tricky. Generative grammars seem to inhabit a space somewhere between Marr's computational and algorithmic level. They do not merely specify which expressions are grammatical and which are not (i.e. a function-in-extension), but specify the rules by which grammatical expressions are formed. However, these rules do not determine an algorithm for performing this computation. I will continue to speak of these theories as computational level theories, but keep in mind that they are slightly more detailed than is typical for theories at this level.

level. The Marrian understanding of cognitive science, however, points to the possibility also of 'vertical' debt collection: theories stated at one level discharging debts accrued at another. Marr's theory is most suited for understanding the second kind of explanatory debt, incurred when one theory relies on a claim that must be verified by another theory. When we explain some observation with reference to a computational theory, we claim that certain computations are implemented in a human brain. That this is so must, ultimately, be established by theories at the algorithmic and implementational level. It is basically a presupposition of contemporary cognitive science that such theories can eventually be provided, and so these debts are not usually viewed as particularly damaging to computational level theories. Cognitive neuroscience is the attempt to discharge these sorts of debts.

The more interesting question is whether there are cases of vertical debt collection of the other sort, wherein apparent exceptions to theories at one level are explained away by theories at another. Perhaps surprisingly, according to a strict interpretation of Marr's theory, they are not. The reason is that Marr's theory posits a strict alignment of levels, so that the behavior (input-output mapping) is the same at each level. Describing the levels involves asking questions about the same system, but at different scales, with reference to different properties, and with different purposes. This means that if some observation seems to be anomalous from the perspective of the computational level, it will likewise be anomalous from the lower levels, which simply show how the higher-level function is computed. For example, if we answer the *what* (as opposed to the *why*) question posed at the computational level by proposing some function from, say, retinal stimulation to representation of object edges, any perceived deviation from this function (e.g. a retinal stimulation that leads to the 'wrong' edge representation) cannot be explained with reference to the algorithmic or implementational level. The reason for this is that these levels provide answers to questions *about* the function computed at the computational level, such as 'what are the representations used by the system in computing this function?', 'what steps does the computing of this function take?', and 'how is this algorithm implemented by a neurobiological system?'. That is, the questions asked at these levels presuppose the correctness of the computational-level description. Thus, on this strict interpretation, deviations from the predicted input-output pattern necessitate revisions to the answers to questions at all levels equally, and so anomalies at one level can't be explained with reference to another.

However, as Marr recognized, this strict interpretation is not quite accurate to scientific practice. In fact, the assumption that the levels are perfectly aligned is an idealization: a certain amount of divergence between levels is permissible. In particular, lower-down levels may well merely approximate the behavior of those higher up. This is clearest in the case of the implementational level at which the stochastic and analog behavior of neurobiology approximates the behavior at the algorithmic level, which is often viewed as both deterministic and discrete. Certain kinds of performance limitations, such as constraints on processing power, may also lead to disparity between levels. When there is such a disparity, the explanatory economy approach will be appropriate. Apparently failed predictions at one level can be explained away with reference to properties of another level. This fact stresses the important of the 'why' aspect of the computational level: when lower levels deviate from the function specified at the computational level, we retain this higher-level specification because

it enables the best explanation of why the system should behave (roughly) this way at all.

In many cases, determining whether a debt ought be paid vertically (by a theory at a different level) or horizontally (by a theory of a different system) may be difficult. For example, failure to grasp a sentence that is predicted to be grammatical could be accounted for either by an algorithmic or implementational theory (vertically) or by a computational theory of a part of the mind other than linguistic competence, such as a parsing theory (horizontally). I have focused on horizontal cases because the relationship between levels of explanation in cognitive science is such a central topic in the philosophy of psychology, while the need to understand the complementarity of theories on the same level but of different targets has been less frequently discussed. However, cognitive science can and should make use of both kinds of debt collection.

9 When debts go unpaid

While the notion of an explanatory economy is helpful in showing how a theory can reject *prima facie* counterexamples without falling into unfalsifiability, it is important to avoid being too permissive. There must be some reasonable prospect of this debt being discharged in order for this move to be legitimate. Without such a prospect, surface appearances are all we have to go on, and *prima facie* counterexamples are just counterexamples. It is crucial that inquiry does not end with the creation of a debt. Instead, this leads to further discussion about the prospects of such a debt being discharged. In fact, the major shifts in the history of generative linguistics can, to a large extent, be viewed as motivated by just such an investigation.

Recall from the previous section the argument from the poverty of the stimulus. This argument aimed to show that various aspects of developed linguistic competence must be determined not by abstraction from linguistic patterns in the environment, but instead by innate (i.e. biological) principles. By positing such innate principles, the problem of language acquisition in the face of the poverty of the stimulus can be solved. However, this solution incurs a debt: these allegedly innate principles must themselves be explained by other theories; biological theories of cognitive development in particular. The central developments in generative grammar can be viewed as attempts to make these debts more suitable for payment by these biological theories.[56]

Early generative theory—transformational grammar—posited language- and construction-specific rules. For just one example, Equi-NP deletion was proposed as the mandatory rule that transforms a construction with the underlying ('deep') structure 'NP_1 V NP_1 VP' (i.e. what are now called 'control structures') by deleting the embedded subject (becoming structures of the form 'NP_1 V VP').[57] Knowledge

[56] I am discussing a very high-level example, in which whole generative approaches are rejected on account of the feeling that the new theory is more likely to have its debts paid. Of course, there are hundreds of examples of lower-level linguistic debates in which all parties agree that incurring a debt would be inappropriate. I focus on the higher-level case because I take it that the worry about the falsifiability of generative grammar is itself a fairly global worry: by denying that certain data are relevant, the generative program is unscientific. I take it that not even the most adamant critics of the program deny that specific generative proposals are sometimes (agreed to be) falsified.

[57] This example is simplified in ways that don't matter for the purposes of this paper.

of this rule was purported to explain why a sentence with the underlying structure 'Edith loves Edith to dance' can only be pronounced after deleting the subject of the embedded clause 'Edith to dance'. This approach to grammar, due to the specificity of such rules, leads very quickly to the positing of an enormous number of rules. Any perceived grammatical relationship between sentences leads to new transformational rules (e.g. passive transformations, question-forming transformations, topicalization transformations, etc.).[58] If, as is often the case, the environmental data underdetermines when such rules are and are not appropriate, these rules must be viewed as innately determined. To pay off the debt that this creates, a biological/developmental story must be, at least plausibly, available for explaining how such rules can be innate.

The move from transformational grammar to Government and Binding theory (GB), and ultimately to the Minimalist Program, is largely motivated by the realization that such debts are unlikely to be paid. Genomics and developmental biology/psychology don't seem to have the tools to account for the development of so many, and such specific, rules as part of developed cognition.[59] Each successive theory therefore attempted to replace the many earlier rules with more general and abstract principles, and this unification led to a reduction in the amount of information assumed to be innate. This greatly increased the prospects of biology paying off this debt. The simpler the grammatical principles posited, and the more generally applicable they were (e.g. Government and Binding theory proposed just one transformation 'move α', for explaining displacement, while the Strong Minimalist Thesis posits the sufficiency of just Merge, which accounts for both displacement and recursive generation), the more plausible it was that this debt could be paid by biological theory. This is especially clearly a motivation in the current Minimalist program, where it is hoped that the properties of, and many of the constraints on, Merge can be explained by very general mathematical properties of complex, emergent systems.[60]

In this way, the idea of explanatory debts can make sense of central developments in linguistic theory. While positing innate structure enabled transformational linguists to provide explanations of competence and development, it did so by incurring an explanatory debt: biology had to make good on its claims of innately determined structure. Rather than this debt simply being accepted, it was actively debated and ultimately deemed unpayable. The response to this was to modify the theory so that the burden placed on other disciplines seemed more reasonable. The push for simpler and more general theories is thus explicable as the attempt to accrue only debts that have a reasonable chance of getting discharged.[61] While still very little is known about

[58] See e.g. Ross (1986) for an early work in which many such specific transformational rules were introduced.

[59] In addition, it was, to say the least, puzzling how to account for such linguistic specificity evolutionarily.

[60] See Chomsky (2007) and Uriagereka (2000) for discussions of this proposal.

[61] This section is probably more accurate as a *rational reconstruction* of the history of generative grammar, rather than a statement of the intentions of its practitioners.

the biological basis for language, the Minimalist Program does seem to provide the best chance for such a unification in the history of generative grammar, and this can be explained by the need to ensure that explanatory debts are at least plausibly payable.

We can briefly note that this approach to the history of linguistics provides a response to Lappin et al.'s (2000) argument that the adoption of the Minimalist Program is an example of an unscientific revolution. They argue argue that linguists who, following Chomsky, shifted from work within the GB tradition to the Minimalist Program are being unscientific in virtue of the fact that their new theory "is in no way superior [to GB] with respect to either predictive capabilities nor explanatory power." (p. 667). Even if this is correct from an *individualistic* perspective, according to which a theory is evaluated purely on its explanations and predictions, it seems false from the the collectivist perspective I have outlined. Even if Minimalism and GB made all the same predictions and explained the same data, one could have an advantage over the other in that the explanatory debts it accrues are more likely to be discharged. I take it that the driving motivation for the Minimalist Program is precisely this: the burden it places on other disciplines is significantly less than that placed on these programs by prior theories, due to the relative complexity of innately determined structure posited by the latter. Of course, this does not show that the Minimalist Program is correct, just that the extent to which a theory cooperates with other theories in other domains is a perfectly 'scientific' reason for adopting such a theory, even if it doesn't lead to enhanced explanatory and predictive power.

Hopefully, this section has made clearer the reciprocal relationship between theories central to the explanatory economy. This points to a more subtle relationship between theory and observation than one finds in standard, individualist, accounts of confirmation. A theory can be confirmed both by explaining/predicting phenomena, and by having phenomena it fails to explain/predict explained by distinct, but compatible, theories. However, if these phenomena cannot be explained by these other theories, then they may be re-evaluated as problematic, perhaps even falsifying, for the original theory. Likewise, the ability of these other theories to discharge the debts of the indebted theory is itself confirmatory for these theories. By paying these debts, these theories indicate that they are externally consistent, a widely accepted theoretical virtue. This produces a sort of 'double-counting' of the confirmatory value of explanations of phenomena that generate debts in other theories: these explanations serve to both increase the explanatory/predictive scope of the theory as well as indicate that this theory is appropriately complementary with other theories. For example, a parsing theory is doubly confirmed by its predictions of the failure to parse multiply center-embedded sentences: as well as the standard mechanism of confirmation by correct prediction, this indicates that a parsing theory and a grammatical theory work well *together*. This shows the importance of the collectivist account of theory confirmation.

Perhaps the most complex relations between evidence and theory involve *failures* to pay off debts created by other theories. When this happens, we have reason to think that this theory is not complementary with the indebted theory.[62] This will be particularly problematic for the theory which may have been expected to discharge

[62] These reasons are not decisive, as the debt could yet be paid off by some third theory, compatible with the first two.

this debt if the indebted theory is highly confirmed, as it suggests that one of these two theories is incorrect. For example, if our best linguistic theories suggest that some unacceptable sentences are grammatical, a parsing theory which does not suggest any difficulty with our interpretation of these sentences is in *prima facie* trouble. This is clearest in the comparative case: if two competing parsing theories are such that only one is able to explain away the apparent anomalies of a well-confirmed grammatical theory, this looks like a pretty good argument in favor of the one that can. The theory that can pay off this debt suggests greater promise of contributing to a collection of theories capable of explaining all the observations, and so is more likely to be true. As all the observations ought be accounted for, a failure of one theory to explain something increases the demand for such an explanation on other theories. Therefore theories that can provide such an explanation are thereby confirmed, indicating their complementarity with other systems, while theories that fail to provide such an account are thereby disconfirmed. Discharging debts is not thus a mere side-effect of successful theories, but one of their essential functions.

10 The explanatory economy and the division of cognitive labor

On the face of it, my proposal seems continuous with a range of work done under the banner of 'The Division of Cognitive Labor', stemming from Kitcher (1990). Strevens (2006, 2003, 2017) has developed Kitcher's picture along a variety of axes, while Zollman (2007) and Weisberg and Muldoon (2009), among others, have developed alternative strategies for modeling the way in which scientific aims are achieved by multiple agents. At a high level of abstraction, these projects are similar to mine in that they aim to understand the ways in which scientific progress is furthered by the differential behavior of multiple agents, and the ways in which the pursuit of certain scientific aims depends on the goals undertaken by others. However, the theory of the explanatory economy I have developed in this paper differs in certain key ways from these other approaches, and even suggests certain difficulties with them. In this final major section I shall outline these differences.

Kitcher and Strevens produce economic models aimed at determining which projects particular scientists ought to undertake, given their perceived chance of success and the social rewards accruing to such success. One can see immediately that this project is quite unlike the one I am engaged in. The explanatory economy is a picture of the epistemic dependence of some scientific theories on others, not of the rational behavior of particular scientists, although the fact that the success of one theory may depend on the subsequent success of another will have important upshots for which projects scientists may want to pursue.

Zollman's approach instead co-opts the tools of network theory, running computer simulations of networks of potential collaborators in order to model the effects that greater openness with respect to the sharing of experimental results has on desirable scientific results such as successfully answering theoretical questions and doing so in a timely fashion. Again, it is clear that Zollman and I simply have different aims. His models assume that each scientific problem (say, the testing of the effectiveness of a certain pharmaceutical) can be solved independently, although collaboration and data-

sharing may influence how frequently and quickly the problems are actually solved. My goal is precisely to understand the ways in which different theories depend essentially on one another for their successes. It may be possible to model the explanatory economy in this network-theoretic way, but it would require a substantial revision of Zollman's models.

Finally, Weisberg and Muldoon's models utilize modified adaptive landscape models, based on those originating in evolutionary theory, to simulate how efficiently a collection of scientists will uncover some set of significant truths. Instead of the simulated landscape representing a distribution of genotypes according to their biological fitness, it represents what Weisberg and Muldoon call 'epistemic significance'. A collection of agents within the model move around this landscape, seeking areas of greatest epistemic significance according to different strategies (for example, some agents are more inclined to follow the paths of other agents, while others are disinclined to do so). Running these simulations can indicate the effects, in terms of how likely it is for the population to uncover the significant truths and how quickly they do so, of having different distributions of strategies in the scientific population.

The central difference between my approach and that of those in the division of cognitive labor tradition is that we are focusing on different kinds of structures. Kitcher, Strevens, Zollman, and Weisberg and Muldoon aim to account for what we might call the *institutional structure* of science. That is, they aim to provide idealized pictures of how actual scientists will or should behave. This is clearest in the case of Kitcher and Strevens, who aim to show why certain behavior is explicable by assuming that scientists are acting so as to increase their scientific prestige. Zollman and Weisberg and Muldoon, on the other hand, aim to model the ways in which different strategies (of e.g. information sharing, or problem-selection) are more or less suited for achieving desirable results such as the timely solution of scientific problems. My focus is instead on the *explanatory structure* of science: the abstract dependencies between the theories and claims put forward by theorists. My approach is thus less directly relevant to explaining and predicting the behavior of scientists. The explanatory economy can show why it was *good* (for a particular theory) that some explanation was provided (e.g. because this explanation discharged a debt incurred by this theory), but it cannot directly explain why some scientists took the time to provide such an explanation.[63]

This said, given that the explanatory economy can show why certain kinds of explanation are needed for the full confirmation of particular theories, by assuming that scientists often aim to produce those explanations that would confirm the theories they pursue, we can leverage the explanatory economy so as to predict/explain scientists' behavior. If you think a theory is correct, one way to contribute to the development and confirmation of this theory is to pay off the debts it has incurred. This motivation clearly has influenced linguistic theorizing. Some of the best research in psycholinguistics is developed precisely to account for apparent anomalies from the perspective of generative grammar. The earlier discussed example of Miller and Chomsky's explanation of the unacceptability of center-embedding is one example,

[63] In this way, my approach is perhaps more akin to work like Mitchell (2003) and Wimsatt (2007), which similarly focus more on the ways in which multiple distinct representations (e.g. theories or models) of complex phenomena can mutually illuminate one another, than it is to the work discussed in this section.

but there are many more.[64] *Prima Facie* it is not clear why these results, as interesting as they are from the perspective of psycholinguistics, are important for theories of *grammar*. The explanatory economy can explain why this is. That is, it can explain why it was beneficial for generative theories of grammar that such explanations were forthcoming. This goes a long way to explaining why these psycholinguistic theories were pursued: these theorists found generative theories of grammar compelling, and aimed to vindicate them by explaining away their apparent anomalies. Of course, whether a theorist is likely to be motivated to discharge a debt accrued to a generative theory of grammar will depend on whether they find such a theory compelling. From the perspective of a skeptic of these approaches, sizable debts may seem like sufficient grounds for rejecting the theory, while from the perspective of an advocate, they will be opportunities for important research.

Despite the crucial difference between accounts of the explanatory and institutional structure of science, I do not view the explanatory economy as irrelevant to standard approaches to the division of cognitive labor. Rather, I think the import of the explanatory economy is that several of these models rest on a picture of scientific progress which is incomplete. In particular, the ways in which scientific progress in one domain is conditionalized on progress elsewhere is absent from the models. Strevens (2003), for example, discusses the 'priority rule', according to which scientific acclaim is attributed exclusively to the scientists who first produce a particular achievement. The explanatory economy, however, points to the fact that this may not be as final as it seems: today's achievements rely on tomorrow's debt collection, and in the absence of this payment what seems like an achievement may not in fact be so. The rewards for such achievements can, in some cases, be re-allocated. This sort of phenomenon is quite common in science: some explanation is produced within a theory which later becomes discredited. In linguistics, many proposed explanations couched in, say, transformational grammar, may have seemed like enduring successes at the time, but are no longer viewed as such given the development of linguistic theory. This may well have an influence on the motivational force of priority in science in general, and in these models in particular. Given that an achievement may be overthrown, scientists may rationally continue work in alternative programs despite the target phenomenon having already been accounted for. Again, this is a common feature of science: consider the wealth of connectionist attempts to account for grammatical phenomena that have been explained by generative linguistics within a symbolic systems framework. Such efforts make sense only on the assumption that the generativists who first explained some linguistic facts will *not* be exclusively rewarded for this work, on account of their explanations ultimately being rejected.

Likewise, my approach may suggest the need for an extra parameter in the models developed in Zollman's more recent work. Zollman (2017) presents a model in which scientists can learn from one another so as to each solve their own problems. The aim of this model is to predict the social conditions which will make scientific collaboration likely and beneficial. Individual scientists are prone to collaborate when doing so

[64] For just a selection of particularly nice examples, see the account of heavy-NP-shift discussed in Staub et al. (2006), Fodor and Inoue (1994) on garden path sentences, Wellwood et al. (2018) on some 'linguistic illusions', and Conroy et al. (2009) for a discussion of children's apparent failure to follow principle B of binding theory.

increases their likelihood of achieving their goals in a timely fashion, without imposing too great a burden. The explanatory economy suggests a further motivation to collaborate: another researcher's work may be needed to underwrite the debts accrued in doing your own. My earlier example of Chomsky's collaboration with George Miller seems like an example of this: Chomsky's linguistic theory seemed unable to account for the unacceptability of multiple center-embeddings, so he collaborated with a psychologist to show how a parsing model could explain these data without modifying the syntactic theory.

Finally, I believe that Weisberg and Muldoon's (W&M) models might provide the best location for incorporating the lessons of the explanatory economy. One way to do this would be to complicate the epistemic landscape and the agents' patterns of moving around it. In particular, while W&M say that they intend the peaks in their epistemic landscapes to be suggestive of the way in which scientific results often depend on one another (pp. 227–228), this dependence is not reflected in the model. The significance of each location is independent of whether other locations have been accessed. One reason W&M make this assumption is that the agents in their model are superb scientists: whenever they move to a particular location they infallibly acquire the epistemic significance of this location. Because of this infallibility, there is no room for the kind of failure that the explanatory economy highlights. A location that seems like it provides a wealth of epistemic significance may do so only on the assumption that the debts engendered in reaching this significance will be discharged. If this doesn't happen, this significance will be illusory. These assumptions (of independence and infallibility) would have to be replaced if one were to incorporate the explanatory economy into a W&M-style model: the epistemic significance of one location could be conditional on the epistemic significance of those locations 'supporting' it (i.e. on a descending path from this location to those locations with lower epistemic significance). If we were also then to incorporate W&M's suggestion (p.233 fn. 6) to distinguish between an agent's exploring the landscape and exploiting it, where the former just means moving around the landscape while the latter means spending some energy extracting the significant truths from their current location, then we may be able to begin modeling the explanatory economy. If we populated the landscape both with genuine peaks of epistemic significance, but also apparent peaks, resting on locations lacking significance (i.e. false assumptions), we could then let agents explore the landscape, again in search of epistemic peaks. An agent confronted with an upwards slope could then be faced with a choice either to climb upwards or start mining the foothills. If they take the former strategy, they increase the speed at which they could reach the top, but they incur the risk that they are climbing a false peak (i.e. that their explanatory debts will not be payed). Discharging epistemic debts could thus be modeled as ensuring that the epistemic significance of higher apparent peaks is suitably supported by the significance of the slopes below. Those at the tops of apparent peaks will thus be dependent on the success of those further down in locating continuous paths of epistemic significance from the floor to the peak. In this way, we could computationally model the ways that different strategies lead to scientific progress. One strategy could involve rushing to the peaks, exploring but not exploiting the path on the way, another could involve more methodically exploiting each step along the way. These would correspond to the approaches of those who aim to produce

the most epistemically significant theories as early as possible (the visionaries), and of those who instead do the dirty work of ensuring that these high-level and apparently most illuminating proposals are consistent with the messy observations (the normal scientists). Each strategy has its benefits and costs. The former involves taking risks for high rewards: impressive theories can be proposed, but they may collapse when it is realized that their debts cannot be discharged. The latter is less susceptible to wasting time developing false theories, but is also less likely to produce the most impressive feats of scientific theorizing. My prediction is that a landscape crafted in this way would be most efficiently populated by a mixed population, featuring some agents looking for the potential areas of greatest significance, and others distinguishing the real from the merely apparent peaks, but one would have to run the models to confirm this.

This has all been highly sketchy, but I hope it indicates ways in which the understanding of the explanatory economy might be able to inform the study of the social/institutional structure of science.

11 Conclusion

In this paper I have shown why the notion of an explanatory economy is a fruitful one, especially as applied to linguistics. In linguistics, the available data are effects of a massively complex system. This makes the task of inferring from data to underlying structure hugely difficult. The Galilean style aims to uncover these underlying structures by treating much of the data set as a noisy interaction effect. This methodology has provoked much consternation as it seems to depart from the empiricist maxim that data are final. I believe that the notion of an explanatory debt can show why this approach is not unscientific as has sometimes been alleged. This methodology should thus appeal to the generativists seeking deep explanatory patterns as well as the more empiricist theorists with greater concern for the quirks and intricacies of used language. With an appropriately collectivist methodology, these approaches can be seen as complementary, taking in one another's difficulties, rather than competing.

Acknowledgements This paper has benefited from conversations with and feedback from many people over the years. In particular, Josh Armstrong, Noam Chomsky, Sam Cumming, Guillermo Del Pinal, Gabe Greenberg, Mark Greenberg, Tim Hunter, Gabby Johnson, David Kaplan, Bill Kowalsky, Kevin Lande, Eliot Michaelson, Shel Smith, Michael Weisberg, as well as two anonymous referees for this journal, have all provided commentary on this work which has led to a much-improved paper.

References

Anderson, J. R. (2009). *How can the human mind occur in the physical universe?*. Oxford: Oxford University Press.

Anderson, J. R. (2013). *The adaptive character of thought*. London: Psychology Press.

Astington, J. W., & Jenkins, J. M. (1999). A longitudinal study of the relation between language and theory-of-mind development. *Developmental psychology*, *35*(5), 1311.

Bošković, Ž. (1998). LF movement and the minimalist program. In *Proceedings of NELS* (vol. 28, pp. 43–57).

Carnie, A. (2013). *Syntax: A generative introduction*. London: Wiley.
Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge: The MIT Press.
Chomsky, N. (1981). *Lectures on government and binding: The Pisa lectures*. Berlin: Walter de Gruyter.
Chomsky, N. (1995). *The minimalist program*. Cambridge: The MIT Press.
Chomsky, N. (2002). *On nature and language*. Cambridge: Cambridge University Press.
Chomsky, N. (2007). Approaching ug from below. In U. Sauerland & H.-M. Gärtner (Eds.), *Interfaces+ recursion= language?* (Vol. 89, pp. 1–30). Berlin: De Gruyter.
Chomsky, N., & McGilvray, J. (2012). *The science of language: Interviews with James McGilvray*. Cambridge: Cambridge University Press.
Churchland, P. M. (1996). *The engine of reason, the seat of the soul: A philosophical journey into the brain*. Cambridge: MIT Press.
Conroy, A., Takahashi, E., Lidz, J., & Phillips, C. (2009). Equal treatment for all antecedents: How children succeed with principle b. *Linguistic Inquiry*, *40*(3), 446–486.
Cooper, R. (1983/2013). *Quantification and syntactic theory* (vol. 21). Berlin: Springer.
Crain, S., & Pietroski, P. (2001). Nature, nurture and universal grammar. *Linguistics and Philosophy*, *24*(2), 139–186.
Crain, S., & Thornton, R. (2000). *Investigations in universal grammar: A guide to experiments on the acquisition of syntax and semantics*. Cambridge: MIT Press.
Culicover, P. W., & Jackendoff, R. (2005). *Simpler syntax*. Cambridge: Oxford University Press.
Cummins, R. (2000). "How does it work?" versus "what are the laws?": Two conceptions of psychological explanation. In F. Keil & R. Wilson (Eds.), *Explanation and cognition* (pp. 117–144). Cambridge: MIT press.
Earman, J. (1992). *Bayes or bust? A critical examination of Bayesian confirmation theory*. Cambridge, MA: MIT Press.
Engdahl, E. (1983). Parasitic gaps. *Linguistics and philosophy*, *6*(1), 5–34.
Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, *32*(5), 429–448.
Ferreira, F., & Patson, N. D. (2007). The 'good enough' approach to language comprehension. *Language and Linguistics Compass*, *1*(1–2), 71–83.
Fodor, J. A. (1981). Some notes on what linguistics is about. In N. Block (Ed.), *Readings in the philosophy of psychology* (Vol. 2, pp. 197–207). Cambridge: Harvard University Press.
Fodor, J. D., & Inoue, A. (1994). The diagnosis and cure of garden paths. *Journal of Psycholinguistic Research*, *23*(5), 407–434.
Freidin, R., & Vergnaud, J.-R. (2001). Exquisite connections: Some remarks on the evolution of linguistic theory. *Lingua*, *111*(9), 639–666.
Ginzburg, J., & Sag, I. (2000). *Interrogative investigations*. Stanford: CSLI publications.
Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.
Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, *15*(2), 135–175.
Huddleston, R. (1994). The contrast between interrogatives and questions. *Journal of Linguistics*, *30*(2), 411–439.
Ibbotson, P., & Tomasello, M. (2016). Evidence rebuts Chomsky's theory of language learning. *Scientific American*, *315*(5)
Jackendoff, R. (1992). *Semantic structures*. Cambridge: MIT Press.
Jackendoff, R. (2010). *Meaning and the lexicon: The parallel architecture 1975–2010*. Oxford: Oxford University Press.
Jackendoff, R. (2018). Representations and rules in language. In B. Huebner (Ed.), *The philosophy of Daniel Dennett*. Oxford: OUP.
Jackendoff, R., & Audring, J. (2019). Relational morphology in the parallel architecture. In J. Audring & F. Masini (Eds.), *The Oxford handbook of morphological theory*. Oxford: Oxford University Press.
Johnson, K. (2009). On failing to capture some (or even all) of what is communicated. In C. Viger & R. J. Stainton (Eds.), *Compositionality, context and semantic values: Essays in honour of Ernie Lepore* (pp. 129–144). Berlin: Springer.
Karmiloff-Smith, A. (1995). *Beyond modularity: A developmental perspective on cognitive science*. Cambridge: MIT Press.
Kayne, R. S. (1994). *The antisymmetry of syntax* (Vol. 25). Cambridge: MIT Press.

Kitcher, P. (1981). Explanatory unification. *Philosophy of Science*, *48*(4), 507–531.
Kitcher, P. (1984). 1953 and all that. a tale of two sciences. *The Philosophical Review*, *93*(3), 335–373.
Kitcher, P. (1990). The division of cognitive labor. *The Journal of Philosophy*, *87*(1), 5–22.
Kitcher, P. (2011). *Science in a democratic society*. Amherst: Prometheus Books.
Kuhn, T. S. (1962/2012). *The structure of scientific revolutions: 50th anniversary edition*. Chicago: University of Chicago Press.
Kuhn, T. S. (2010). Objectivity, value judgment, and theory choice. In A. Bird & J. Ladyman (Eds.), *Arguing about science* (pp. 74–86). London: Routledge.
Labov, W. (1971). The notion of 'system' in creole studies. In D. Hymes (Ed.), *Pidginization and creolization of languages* (pp. 447–472). Cambridge: Cambridge University Press.
Labov, W. (1975). Empirical foundations of linguistic theory. In R. Austerlitz (Ed.), *The scope of American linguistics* (pp. 77–133). Berlin: De Gruyter.
Lakatos, I. (1976). Falsification and the methodology of scientific research programmes. In S. Harding (Ed.), *Can theories be refuted?: Essays on the Duhem-Quine Thesis* (pp. 205–259). Berlin: Springer.
Lappin, S., Levine, R. D., & Johnson, D. E. (2000). The structure of unscientific revolutions. *Natural Language & Linguistic Theory* (pp. 665–671).
Longino, H. E. (1995). Gender, politics, and the theoretical virtues. *Synthese*, *104*(3), 383–397.
Manning, C. D. (2003). Probabilistic syntax. In J. Hay, S. Jannedy, & R. Bod (Eds.), *Probabilistic linguistics* (pp. 289–341). Cambridge: MIT Press.
Marcus, G. F. (2001). *The algebraic mind: Integrating connectionism and cognitive science*. Cambridge: MIT Press.
Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Cambridge: MIT Press.
May, R. (1985). *Logical form: Its structure and derivation*. Cambridge: MIT Press.
Miller, G. A., & Chomsky, N. (1963). Finitary models of language users. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology*. London: Wiley.
Mitchell, S. D. (2003). *Biological complexity and integrative pluralism*. Cambridge: Cambridge University Press.
Newell, A. (1982). The knowledge level. *Artificial intelligence*, *18*(1), 87–127.
Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175.
Phillips, C. (2004). Linguistics and linking problems. In S. F. Warren & M. Rice (Eds.), *Developmental language disorders: From phenotypes to etiologies* (pp. 241–287). Trenton: Lawrence Erlbaum Associates.
Piattelli-Palmarini, M., Uriagereka, J., & Salaburu, P. (2009). *Of minds and language: A dialogue with Noam Chomsky in the Basque Country*. Oxford: Oxford University Press.
Popper, K. (1959/2002). *The logic of scientific discovery*. London: Routledge.
Purver, M. R. J. (2004). *The theory and use of clarification requests in dialogue*. Ph.D. thesis, University of London.
Rizzi, L. (1996). Residual verb second and the wh-criterion. In A. Belletti & L. Rizzi (Eds.), *Parameters and Functional Heads: Essays in Comparative Syntax* (pp. 63–90). Oxford: Oxford University Press.
Rosenberg, J. F. (1988). About competence and performance. *Philosophical Papers*, *17*(1), 33–49.
Ross, J. R. (1986). *Infinite syntax!*. Norwood: Ablex.
Scriven, M. (1959). Explanation and prediction in evolutionary theory. *Science*, *130*(3374), 477–482.
Shagrir, O. (2010). Marr on computational-level theories. *Philosophy of Science*, *77*(4), 477–500.
Sober, E. (1984). *The nature of selection: Evolutionary theory in philosophical focus*. Chicago: University of Chicago Press.
Sobin, N. (1990). On the syntax of english echo questions. *Lingua*, *81*(2–3), 141–167.
Spelke, E. S. (2003). What makes us smart? core knowledge and natural language. In S. Goldin-Meadow & D. Gentner (Eds.), *Language in mind: Advances in the study of language and thought* (pp. 277–311). Cambridge: MIT Press.
Stabler, E. P. (1991). Avoid the pedestrian's paradox. In S. Abney & R. C. Berwick (Eds.), *Principle-based parsing: Computation and psycholinguistics* (pp. 199–237). Berlin: Springer.
Staub, A., Clifton, C, Jr., & Frazier, L. (2006). Heavy np shift is the parser's last resort: Evidence from eye movements. *Journal of Memory and Language*, *54*(3), 389–406.
Strevens, M. (2003). The role of the priority rule in science. *The Journal of Philosophy*, *100*(2), 55–79.

Strevens, M. (2006). The role of the matthew effect in science. *Studies in History and Philosophy of Science Part A*, *37*(2), 159–170.
Strevens, M. (2017). Scientific sharing: Communism and the social contract. In T. Boyer-Kassem, C. Mayo-Wilson, & M. Weisberg (Eds.), *Scientific collaboration and collective knowledge*. Cambridge: OUP.
Tomasello, M. (2009). *Constructing a language*. Cambridge: Harvard University Press.
Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences*, *28*(5), 675–691.
Townsend, D. J., & Bever, T. G. (2001). *Sentence comprehension: The integration of habits and rules*. Cambridge: MIT Press.
Trinh, T. H. (2011). *Edges and linearization*. Ph.D. thesis, Massachusetts Institute of Technology.
Uriagereka, J. (2000). *Rhyme and reason: An introduction to minimalist syntax*. Cambridge: MIT Press.
Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, *44*(1), 92–107.
Weisberg, M. (2006). Robustness analysis. *Philosophy of Science*, *73*(5), 730–742.
Weisberg, M., & Muldoon, R. (2009). Epistemic landscapes and the division of cognitive labor. *Philosophy of Science*, *76*(2), 225–252.
Wellwood, A., Pancheva, R., Hacquard, V., & Phillips, C. (2018). The anatomy of a comparative illusion. *Journal of Semantics*, *35*(3), 543–583.
Wilson, B., & Peters, A. M. (1988). What are you cookin' on a hot?: Movement constraints in the speech of a three-year-old blind child. *Language* (pp. 249–273).
Wimsatt, W. C. (2007). *Re-engineering philosophy for limited beings: Piecewise approximations to reality*. Cambridge: Harvard University Press.
Zollman, K. (2017). Learning to collaborate. In T. Boyer-Kassem, C. Mayo-Wilson, & M. Weisberg (Eds.), *Scientific collaboration and collective knowledge*. Oxford: Oxford University Press.
Zollman, K. J. (2007). The communication structure of epistemic communities. *Philosophy of Science*, *74*(5), 574–587.
Zwicky, A. M., & Pullum, G. K. (1987). Plain morphology and expressive morphology. In *Annual meeting of the Berkeley Linguistics Society* (vol. 13, pp. 330–340).