Language, Cognition and Neuroscience, 2015
Vol. 30, No. 4, 357–366, http://dx.doi.org/10.1080/23273798.2014.980750

Routledge
Taylor & Francis Group

Towards a computational(ist) neurobiology of language: *correlational*, *integrated* and *explanatory* neurolinguistics

David Embick[a]* and David Poeppel[b,c]

[a]*Department of Linguistics, University of Pennsylvania, 619 Williams Hall, Philadelphia, PA 19104-6305, USA;* [b]*Department of Psychology, New York University, 6 Washington Place, New York, NY 10003, USA;* [c]*Max-Planck-Institute (MPIEA), Grüneburgweg 14, 60322 Frankfurt, Germany*

We outline what an integrated approach to language research that connects experimental, theoretical and neurobiological (NB) domains of inquiry would look like and ask to what extent unification is possible across the domains. At the centre of the programme is the idea that computational/representational (CR) theories of language must be used to investigate its NB foundations. We consider different ways in which CR and NB might be connected. These are (1) a *correlational* way, in which NB computation is correlated with the CR theory; (2) an *integrated* way, in which NB data provide crucial evidence for choosing among CR theories; and (3) an *explanatory* way, in which properties of NB explain why a CR theory is the way it is. We examine various questions concerning the prospects for explanatory connections in particular, including to what extent it makes sense to say that NB could be *specialised* for particular computations.

Keywords: computation; linguistic theory; unification; explanation; reductionism; levels

1. The most general question confronting current research in cognitive neuroscience is how increasingly detailed neuroanatomical and imaging work can be employed in the development of theories in which there are *explanatory* connections between cognition and neurobiology (NB). Recent technological advances constitute great improvements over earlier ways of looking at brain function. At the same time, though, it is not clear that these technological successes have been matched by the conceptual advances that are required for the development of truly integrated or *unified* theories that link cognition and NB. Thus, although cognitive theories and NB theories are advancing in their own terms, there are few (if any) substantive *linking hypotheses* connecting these domains.

We take the questions about how to unify cognitive and NB lines of inquiry to be completely general to the cognitive neurosciences. Under ideal circumstances – i.e. with sufficient time and resources – it would be worth looking in detail for linking hypotheses throughout the entire 'human cognome' (the comprehensive 'parts list' of the human mind), in order to determine what progress towards unification has been made after 25 years of intensive neuroimaging.[1] In this paper, we frame some of the foundational issues confronting unification with reference to our narrow focus, language.[2] It is fair to say that the perspective we develop here differs markedly from most research in cognitive neuroscience of language, where it is typically assumed that neural data have 'epistemological priority' over data from linguistics or psycholinguistics. We argue instead that computational-representational (CR) theories that are the foundation of language research invite not reduction to the biological infrastructure that is described by the neurosciences, but rather conceptual change and ultimately unification – subsequent to the identification of linking hypotheses between these domains of study. In short, we support a more 'muscular' linguistics and cognitive science.

2. We assume that *language* comprises a set of representations (e.g. 'morpheme') and computations (e.g. 'concatenation') whose formal properties are the object of research in (psycho)linguistic theory as currently practiced. We assume, moreover, a view of language that takes it to be part of the natural sciences, in which questions of biology, in general, and neurobiology, in particular, play a prominent role – in principle, if not in practice.[3]

Our discussion is directed at a specific question that is presumably the centre of the 'cognitive neuroscience of language' (and which, in terms of our first section, could be generalised to cognitive neuroscience as a whole): *Can discoveries about the structure and functional organization of the brain* ***explain*** *properties of the computations and representations that constitute language? That is,*

*Corresponding author. Email: embick@ling.upenn.edu

Our goal in this paper is to outline a framework for understanding connections between computational–representational and neurobiological theories of language. For a variety of reasons (both of principle and of practice) we have been unable to discuss in-depth a wide range of actual examples from current language work. Our hope is that we will be able to present a number of such case studies in a companion piece to this article.

under what conditions can it be said that there is an explanatory relationship between some aspect of neurobiology and some aspect of linguistic computation?

Our goal here, as in the case of earlier forays into these themes (Poeppel, 2012; Poeppel & Embick, 2005), is to introduce a particular set of questions into discussion of language and the brain and to outline a research programme that attempts to address them. We do this with particular emphasis on some possible forms that answers to these questions might take, in the hope that this will provide a concrete foundation for further development.

3. Since our goal is to address the potential for unifying two lines of inquiry, from the cognitive and NB domains, some initial assumptions concerning the nature of each need to be clarified. It is convenient to develop some of our main themes with reference to Marr's (1982) distinctions between the computational, algorithmic and implementational levels of analysis:

> Computational theory: What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?
>
> Representation and algorithm: How can this computational theory be carried out? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?
>
> Hardware implementation: How can the representation and algorithm be realized physically?

Current research on language in the cognitive domain concentrates on the first two levels; over the last 60 years or so of research, much progress has been made, especially considering the immense complexity of language and its connections with other systems. The theories that have been developed make claims both about how language is structured as a formal system, a *grammar*, and there are also theories of how grammars are used by speakers in real time in production and in comprehension. We use *(psycho)linguistic theory* as a cover term for these lines of research on language, with no distinction being made at present between 'theoretical linguistics' and 'psycholinguistics' (the typical categories for the former and latter types of theories mentioned above). When it becomes relevant later in the paper (Section 12), we make some additional comments concerning distinctions between psycholinguistics and linguistic theory that are potentially important for our explanatory goals.

Theories of the (psycho)linguistic type, which make specific claims about the *computations* and *representations* that constitute grammar and aspects of language use, fall under the general heading of '*Computational-Representational*' (*CR*) theories (for this terminology, see the papers collected in Chomsky (2000)). Here we sometimes use the term *computations* as a cover for both computations and representations.

The second line of research, implementational in character, comes from *NB*: the study of the structure and function of the brain. Under this general heading we include any approach that makes claims about how brain structures and different forms of brain activity underpin perception and cognition. In practice, theories on the NB side differ greatly along various dimensions; *e.g.* in terms of the size of the objects they concentrate on (small = channels, neurons; large = populations, areas, *etc.*), or in terms of the methodologies that they employ to study the brain (different types of invasive and non-invasive imaging, with different dependent variables). For present purposes, we look beyond these differences and treat all brain-centred theories of cognition, ranging from biophysical models to systems – and cognitive – neuroscience, under the heading *NB*.

When we speak of potentially unifying CR and NB theories, we do so with a very specific understanding of the overall goal of research in cognitive neuroscience: the goal is to understand why language has the cognitive – that is to say, computational and representational – properties that it has. This stance might look like we are privileging CR, since clearly a CR type of theory looks like it is better adapted to explaining cognition than an NB theory. This is not our intention, though; it is not our intention to simply assert that CR theories of language, which are by definition interested in cognition, are prima facie better at explaining the cognitive properties of language than NB theories. This point is obvious, but not very instructive. Our goal is to highlight the idea that some crucial explanatory force in understanding language will come from understanding neurobiology. Thus, we do not see any point to attempting to achieve *reduction* of CR to NB, or vice versa (whatever that would mean). Rather, our goal is to outline a framework for unified inquiry into language, in which CR and NB are investigated together and on equal epistemic footing.

4. There are two major problems that can be identified when we try to link CR and NB theories (Poeppel & Embick, 2005). The first is that CR and NB theories have different types of primitives, i.e. distinct ontologies, making any attempts at directly linking the two domains prima facie problematic, if not incoherent. This problem is referred to as the Ontological Incommensurability Problem (OIP). To illustrate, consider that the CR theories and NB theories are each advancing in their own terms, such that each has developed its own inventory of primitive objects and elementary operations on these objects, along the lines listed in Table 1.

Table 1. Examples of hypothesised primitive objects/operations.

	Linguistics	Neuroscience
Objects	Distinctive feature	Dendrite/spine
	Timing slot	Neuron
	Morpheme	Cortical microcircuit
	Phrase	Cortical column
Operations	Feature spreading	Long term potentiation (LTP)
	Merge	Oscillation
	Concatenation	Adaptation
	Semantic composition	synchronisation

The problem is that one cannot simply 'draw lines' between the categories provided by each domain and expect such an attempt at 'alignment' to withstand any serious scrutiny. For example, the claim that the object *morpheme* in the CR theory corresponds to the object *neuron* in the NB theory is a non-starter; it is not even wrong. But crucially, without any straightforward way of aligning and connecting the inventories that the two types of theories identify, it is not possible to speak of any sort of unification. The 'mappings' between domains are at best correlational. Such correlations are, to be sure, positive results. Identifying the 'neural correlates' of any perceptual or cognitive function is often the key research goal (consider, e.g. the vigorous search for the neural correlates of consciousness, NCC). For example, neuronal activation patterns that are reliably observed in response to morphological decomposition (e.g. Fiorentino & Poeppel, 2007; Marslen-Wilson & Tyler, 2007) inform us about what brain regions or neural response patterns correlate with the manipulation of morphological information; however, such correlative information does not constitute an *explanation* of why morphemes are represented and processed the way they are in light of the neural data. That desideratum is much harder to satisfy.

The OIP is a problem in principle, not in practice; the question of how computations and representations relate to NB structure is general to domains of research comprising contemporary cognitive neuroscience, classical cognitive science, the philosophy of mind and a number of adjacent areas of inquiry. The reason the alignment merits an answer, of course, is (1) that the primitives stipulated in the CR and NB domains are highly successful empirically in their own right and (2) that there is no principled reason to assign epistemological primacy to either one or the other set of hypothesised primitives. The (almost reflexive) reductionist stance of mapping putative CR primitives to NB structures as 'more basic' is unmotivated and in our view constitutes an impediment to progress on CR/NB connections. At the same time, in truly integrated theories, NB data are in principle just as important as other kinds of data for understanding what types of computations and representations are involved in language (see below).[4]

A second problem for the development of CR/NB linking hypotheses is practical (and therefore, tractable) and derives in part from the relative stages of development of (psycho)linguistic theory, on the one hand, versus theories of language in the brain, on the other. In the way they treat details of linguistic structure in different parts of language, CR theories are fine-grained. However, current theories of linguistic representations and computations *as realised in the brain* are relatively coarse-grained. What we mean by this is that while CR theories of language make predictions about the details of objects in many different domains of language – e.g. about the specific sub-computations that make up larger domains like 'phonology', 'morphology', 'syntax', 'semantics' – many neurolinguistic theories are at this point in time primarily examining differences between coarser subdivisions: e.g. where 'syntax' as opposed to 'phonology' is in the brain. The difference between the types of questions being posed in the CR and in the NB theories makes unification difficult because current work in the CR paradigms makes predictions about distinctions that are at present not typically even considered in the NB domain, and NB studies might be looking for a cognitive 'object' like *syntax* that is actually internally complex, in a way that renders questions like 'where is syntax?' potentially misleading oversimplifications. In terms of Poeppel and Embick (2005), this set of differences in detail of analysis is collectively referred to as the Granularity Mismatch Problem (GMP).

In highlighting the GMP, we are not asserting that neuroscience itself is coarse-grained; to the contrary, spectacular progress has been made in identifying the brain's structures and operations. But these accounts of neural structure and function do not connect with (or map to) the objects employed in CR theories: it is difficult to establish CR/NB linking hypotheses because in general, the study of how the brain computes what it computes in language is at present too coarse to link up meaningfully with the distinctions made on the CR side. Our view is that this situation can be fixed by investigating the brain in terms of CR objects like those in Table 1, which have established value in understanding the structure of language. More precisely: *it is only when computational primitives are examined that possible links between CR and NB could be forged*; current research programmes typically ask about syntax, phonology, semantics, etc., as if they were monolithic entities, in a way that – we believe – constitutes an obstacle to serious advances on the interdisciplinary unification problem.

5. A path towards a research agenda that places possible CR/NB connections at the centre begins with the following two questions:

(Q1) *Basic Question*: How does the brain execute the different computations that make up language?

(Q2) *Advanced Question*: Is the fact that human language is made up of certain computations (and not others) explained by the fact that these computations are executed in neurobiological structures that have certain properties (and not others)?

Regarding Q1, we have argued that the development of CR theory is an absolutely essential step towards understanding the NB of language because CR theory provides the primitive computations that could be meaningfully connected with the NB ontology. Adopting such a view, it appears that there are at present three possible ways in which CR and NB could be found to relate to each other:

Correlational neurolinguistics: CR theories of language are used to investigate the NB foundations of language. Knowledge of how the brain computes is gained by capitalising on CR knowledge of language.

Integrated neurolinguistics: CR neurolinguistics *plus* the NB perspective provides crucial evidence that adjudicates among different CR theories. That is, brain data enrich our understanding of language at the CR level.

Explanatory neurolinguistics: (Correlational + Integrated neurolinguistics) *plus* something about NB structure/function explains *why* the CR theory of language involves particular computations and representations (and not others).

Question Q2 above implicates *specialisation*: it asks whether the reason that we find particular computations in language is explained by properties of the neurobiological structures in which those computations are represented and executed in the brain. It relates directly to *explanatory* neurolinguistics.

These three types of CR/NB connection are examined in Sections 6–9. After this, we begin a more detailed look at Q2, with a focus on the idea that NB structures might be specialised in CR-relevant ways.

6. *Correlational neurolinguistics* in essence is the idea that computational theories of language can be used profitably as a basis for exploring the brain, and that using CR theories in this way will tell us how the brain represents and computes language, or what NB structures and response patterns *correlate* with the representations and computations posited in the CR theory.

Although this is the simplest of the three types of CR/NB connection that we posit, a theory of correlational neurolinguistics would be a huge achievement. In order to produce a working theory of how the brain computes, we would have to answer questions about encoding of information in the brain, the nature of the operations that apply to such representations, and so on, that are at the frontiers of research in essentially all domains of language. A survey of where the field lies with respect to correlational neurolinguistics is beyond the scope of this paper; suffice it to say that at present there is no clear idea of how the brain represents and computes *any* of the computations that are part of language (*cf.* Gallistel & King, 2009). This does not mean that no progress has been made – in fact, a number of highly important and suggestive correlates of different aspects of linguistic computation have been identified (see section 8 below for some illustrations). However, such correlations are only a preliminary step towards questions about encoding of representations and computations that would be the core substance of a theory of language in the brain.

7. Under the heading of *integrated neurolinguistics* we have in mind situations in which information derived from the types of variables that are examined in neuroscience prove decisive in selecting from competing CR-theoretical options. In particular, we hypothesise that there might be scenarios in which the types of information made available by 'typical' methods employed in (psycho)linguistics (i.e. patterns of linguistic phenomena in the languages of the world, or behavioural data of different types, such as lexical decision times, judgements or eye movements) underdetermine a choice among different and competing CR theories. It is to be hoped that data from the NB side would provide the decisive evidence that one of the theoretical options is preferred. If this were to happen, the CR and NB dimensions would be truly unified, since the question of which theory is correct *requires* information from both sides so that (in effect), CR and NB inquiry turn out to be simply different methodologies for investigating (and presumably answering) the same theoretical questions.[5]

In practice, illustrating this type of CR/NB interaction further requires a closer look at the CR side itself, since, in current work, it is unclear to what extent the *theoretical* linguistic and *psycho*linguistic modes of inquiry are actually unified in this sense. An active topic of ongoing research in CR theories is where the dividing lines should be drawn between the computational aspects of language and the representation/algorithm. Questions of this type have a long history, intersecting with key themes, such as the (controversial) distinction between *performance* and *competence* (Chomsky 1965) that were part of the development of linguistic theory in the middle of the last century. Roughly speaking, in the current research, the 'real-time' properties of language that implicate specific algorithms, etc., are pursued under the heading of *psycholinguistics*, whereas the computational characterisation is the domain of *theoretical linguistics*; at least, according to one view.

It is difficult to find examples in the context of CR theory in which competing theoretical options about the structure of language have been resolved by data available only from behavioural studies of language processing (or for that matter, from the study of language acquisition).[6]

Rather, in much of the work that links theoretical and psycholinguistic work directly, the connections seem to be of the type that we point to under the label of *CR neurolinguistics* above, where the representations and categories employed in the theoretical analysis are used to probe and interpret matters of psycholinguistic relevance (Lewis & Phillips, in press; Phillips & Lau, 2004).

In part, the relative disconnect in this domain is found because theories of grammar and psycholinguistic theories are often treated as if they were directed at different objects of inquiry – internalised systems of grammatical competence in the former case, versus systems that employ grammars in real time in the latter case – in a way that (more or less directly) reflects a difference between the *computational* and *algorithmic* levels of analysis in Marr's sense.

At the same time, there are at least some instances in which the concerns of theoretical linguistics and psycholinguistics overlap directly so that questions about unification can be raised. For example, it appears at present that distributional facts about languages come up short in some important (and relatively well-studied) parts of grammar. The analysis of (certain) 'irregular' verb forms in English and other languages, the topic of an intense discussion from the late 1980s onwards under the heading of the 'past tense debate' (see e.g. Marslen-Wilson & Tyler, 1998), appears to be of this type. Using, for illustration, the verb *sing* and its past tense form *sang*, there are two main ways of analysing the past tense form: one that says that it is derived by (morpho)phonological processes that operate on an underlying representation that is shared by *sing*, and another type of approach that relies more on memory, and holds that *sang* is stored as an unanalysed whole. Essentially all theories of grammar hold that both phonological alternations and memorised alternants are required in the theory to some degree. Thus, either analysis could in principle extend to *sing*/*sang* and related alternations. Importantly, while phonological alternations and alternations involving memorised forms appear to operate under different structural conditions (different *locality conditions*, in linguistic parlance), these criteria do not appear to be decisive in determining the status ('memorised' or 'derived by rule') of alternations like that seen in *sing/sang* (cf. Embick 2010a, 2010b). In this type of situation, it appears that the crucial information adjudicating between the competing CR theories must come not from distributional patterns in languages, but from other domains – i.e. from the types of variables that are available in psycho- and neurolinguistic inquiry.[7] In situations where decisive evidence comes from the NB side, there would be integrated neurolinguistics in our sense.

8. In practice, the boundaries between correlational neurolinguistics and integrated neurolinguistics are somewhat fluid. The reason for this is that different research programmes (and different researchers) have different views on which CR theories of language are correct and different intuitions about the relative contributions of CR-based and NB-based findings to the overall theory of language and brain.[8]

A specific example of the different views that can be taken on CR/NB connections will help to illustrate these points. Our example is drawn from the sound side of language, phonetics and phonology. A foundational question in this domain, which has been at the centre of research for more than a hundred years, concerns *the elementary building blocks of speech sounds*. A typical answer to this question from outside of linguistic theory might be 'phonemes' (or 'segments'). That is to say, the atomic level of description is assumed to be equivalent to a segment of sound about which one might have intuitions (say the /k/ sound at the beginning of the word *cat*) or which might correspond, if loosely, to letters in an alphabetic script. However, it has been well-established in language research (based on theoretical considerations concerning the 'moving parts' of phonological alternations; studies of historical change and synchronic variation; cross-linguistic research; developmental studies; aphasia research) that there is a critical layer beneath the segment or phoneme: ***distinctive features***. Since the work of Trubetzkoy (1939), Jakobson, Fant, and Halle (1951), and many subsequent analyses (e.g. Chomsky & Halle, 1968; Halle, 1959), there is a consensus that contrasts between phonemes can be best and most effectively explained in terms of *properties* of segments – features – rather than treating segments as atomic. Individual speech sounds are thus composed of bundles of features, a view that has wide support, although the details of some parts of the theory (e.g. the number and specification of features) remain under investigation. The features, which define natural classes, typically not only refer to articulatory primitives (e.g. 'place of articulation' and 'voicing') but also have acoustic interpretations (Stevens, 2002). In linguistic research, the notion that there is a featural level of organisation beneath the segment is more accepted than the status and value of a segment itself.

These key insights from linguistic research can be used in neuroscience experiments that investigate how the brain represents speech sounds. The distinction between features versus phonemes (or segments) allows us to understand different brain responses and what underlying operations and representations they likely reflect. There exists a growing body of experimental work that shows at what stage of processing neural responses demonstrate a compelling sensitivity to the featural organisation of speech. For example, electrophysiological studies using magnetoencephalography (MEG) and electrocorticography (ECog) have shown that responses to speech sounds between 100 ms and 200 ms are finely tuned to featural

information (e.g. Mesgarani, Cheung, Johnson, & Chang, 2014; Obleser, Lahiri, & Eulitz, 2004; Phillips et al., 2000; Scharinger, Merickel, Riley, & Idsardi, 2011). To be sure, there is a much larger neurolinguistic literature investigating how phonemes are mapped (largely in spatial terms, i.e. 'phoneme-o-topy'). The challenge for correlational neurolinguistics and integrated neurolinguistics is to test whether the well-supported CR theories like those that build on feature theory provide a better way for aligning with neural data than CR theories with other ontologies (e.g., with atomic phonemes) do. To the extent that this question is experimentally resolved, it is then possible to claim that there is progress in the joint understanding in the CR/NB properties of language.

This example from phonological theory illustrates the ways in which objects from a CR theory (in this example, phonological features, as opposed to phonemes) can be employed to investigate how the brain represents speech sounds. Taken in this way, it is an example of correlational neurolinguistics as defined above.

At the same time, this example could also be viewed from the perspective of integrated neurolinguistics. In a way that perhaps depends on the assumptions motivating a particular research programme, the theoretically motivated arguments for features over phonemes that are cited above could be deemed to be inconclusive. (We do not hold this view, but there are many who do, especially in more neurologically oriented work). In this (for us hypothetical) context, a scenario in which brain data could better be understood in terms of a feature-based model, as opposed to a phoneme-based model, would constitute an instance in which brain data provide crucial evidence that adjudicates between competing CR theories – integrated neurolinguistics, as we have defined it above.

9. The intuition behind *explanatory neurolinguistics* extends beyond correlational and integrated neurolinguistics. The idea is that there will be situations in which NB and CR *must* be pursued together because the NB part of the theory *explains* why it is that particular CR computations and representations are employed. An important assumption here, which derives from the CR study of language, is that the core computations in language are relatively limited in number; although this observation takes different forms in different subdomains of language, it can be asked across the board why *those particular* computations and representations – and not others – are found.

A key question for explanatory neurolinguistics is what it could mean for part of the NB ontology to be *specialised* for a set of computations. It seems reasonable to assume that when NB is specialised for a particular type of CR (and not 'general') NB could truly explain why CR is the way it is. In language (and, for that matter, in other domains), it is extremely difficult at present to conceive of the forms this kind of specialisation might take. In part, this is because the ontological incommensurability problem (OIP) has not been solved for language. More generally, it is because the question of how biological systems are specialised for different computations or algorithms has not been answered in more than a few cases. For this reason, we will try to summarise an example from another domain in which CR/NB explanation is at issue.

Our example is drawn from one of the major success stories of biology, in general, and neuroethology, in particular. It centres on our understanding of how temporal information is encoded and represented in animals to construct an internal model of space that facilitates essential behavioural requirements, such as sound localisation in the context of predator/prey interactions. In that context, the detection of interaural time differences (ITDs) has received particular attention and has been the subject of numerous studies using psychophysical, neurophysiological and computational approaches. The reason we are interested in these phenomena has to do with the fact that this research constitutes a successful example of the Marr-inspired research programme. It has been possible to demonstrate a set of neural circuits in the barn owl (avian example) and the gerbil and guinea pig (mammalian example) that execute highly specific and different operations. In the barn owl, the computational model for sound localisation suggested originally by Jeffress (1948), which consists of a coordination of delay lines and coincidence detectors, has been discovered in the NB. Cells in the nucleus laminaris receive and detect delay-line coded input from the more peripheral nucleus magnocellularis. Interestingly, the high-level computational theory of the task – sound localisation based on temporal cues – turns out to be subserved by different algorithms in the avian versus mammalian auditory brainstem circuitry, and the specific cellular implementation reflects a specialisation for one type of algorithm or the other. We take this research on sound localisation to show not only that circuits execute *particular* computations, but also that there are highly constrained and, we believe, explanatory relationship between properties of the circuit, the algorithms carried out by that circuit and the appropriate high-level computational task.

In what sense do we believe that there is genuine specialisation and genuine explanation in the example we raise here? Our perspective derives from two critical observations. First, considering the computational level of description of the problem (how to use auditory temporal information in signals to calculate positional information), it is important to appreciate that different algorithms are appropriate matches to the task at hand. That is, the computational analysis alone is underspecified with respect to the underlying algorithm. In the case of

interaural time information, there are at least two, but probably more, algorithms that successfully use ITD to derive target location data. However, when mapping from algorithm to the implementational level of description, the strictures are more apparent. Demonstrably, there is one type of circuit that employs a Jeffress-type algorithm to use ITD. However, unlike in the avian case, the mammalian system employs a slightly different, phase-based calculation and, crucially, the neuronal circuitry underlying that computation is a bit different (see Grothe, 2003 for an excellent review).

In summary, this example provides a concrete instance of CR/NB specialisation. Moreover, the finding that different algorithms are used for the same computational problem, and are computed in different NB structures, is an important one (see Section 14 below).

In language it is not clear what form explanatory neurolinguistics might take. For this reason, we will concentrate in the sections to come on refining the questions at issue.

10. The fact that it is hard to outline examples of specialisation could be an indication that we have posed our defining questions badly, or that we have pursued a path that is likely to lead nowhere. We think, though, that abandoning the line of inquiry surrounding explanatory CR/NB connections would be a serious error. *If unification is the ultimate goal of pursuing research in cognitive neuroscience – and we believe that it is – then there must be an articulated conceptual framework for understanding how unification is to proceed, and what forms it might ultimately take*. Thus, while our comments here are programmatic, we hope they provide the basis for more sustained investigations of what NB specialisation for CR computation might look like.

In the rest of this paper we develop two questions about the nature of specialisation:

> *Specialisation question 1*: Are there particular levels of NB organisation that are to be privileged as candidates for CR specialisation?
>
> *Specialisation question 2*: Are there particular parts of the CR theory that are more likely to be candidates for explanatory neurolinguistic explanation than others?

The first question directly implicates different research programmes in contemporary cognitive neuroscience, which appear to embody distinct hypotheses about what level(s) of NB organisation could potentially be used to explain CR (or perhaps more precisely, they differ with respect to what levels of NB organisation they privilege).

The second question points to the idea that some CR properties might be subject to explanatory neurolinguistic connections, whereas other aspects of CR might not be. It is conceivable that we might discover, for example, that certain facets of language are the way they are not because of the way the brain is organised but because of other factors; *e.g.* 'general' principles governing computation in biological systems, the interaction with other language-external factors (*e.g.* memory and processing capacity), and so on (see, *e.g.* Chomsky, 2005 for pertinent discussion).

11. The question of what level of NB structure should be targeted for specialisation is very broad in scope. One way of looking at many of the most influential research programmes in the field is in terms of what objects they are betting on for being of computational interest.

For instance, research programmes centred on neural networks, which examine (models of) neurons (and their connections), seem to be making the claim that it is at a (relatively) small level of NB organisation that properties of computational interest are to be sought. The positive and negative points of these systems continue to be examined in an active literature in the CR domain. On the NB side, such models have apparent prima facie credibility because of their putative relation to 'known' objects in NB (neurons). It is relevant for our purposes to note that neurons are only one object (or level of analysis) among many to investigate as far as CR/NB specialisation goes, and that research programmes that restrict themselves to neurons do so on the basis of a research intuition, not on the basis of anything else.

Looking at the other end of the size scale moves the discussion to one of the main themes in cognitive neuroscience as currently practiced, particularly in the domain of neuroimaging (which has inherited much from earlier neuropsychological models): this is the idea that anatomically defined areas (e.g. Brodmann areas or something along those lines) are an appropriate level of NB structure to look at for interesting computational properties. This kind of reasoning (which is often implicit) is what is behind the idea that certain 'areas' of the brain (e.g. Broca's area), 'do' certain parts of language (e.g. 'syntax'). For reasons that we have discussed in other work (Embick & Poeppel, 2005; Poeppel & Embick, 2005), we believe that there are serious limitations to this kind of 'localizationist' research programme.[9]

Our view is that there is not necessarily one particular size of NB organisation that we should be committed to a priori. This claim embodies a kind of 'computationalist' perspective, of a type that has been expressed by Gallistel and King (2009), among others. According to their formulation, it is necessary to '… draw … architectural conclusions from computational commitments', not vice versa; i.e. rather than restricting oneself in advance to one part of the NB ontology from the beginning, like neurons, meaningful connections should be sought in a way that is directed by the CR problem that is addressed and in ways

that could implicate larger or perhaps even smaller NB structure.[10]

With respect to larger NB structures, one intuition that we are sympathetic to is the idea that CR/NB connections should be sought at the level of the neural *circuit*. Informally, our view is that the goal of research in cognitive neuroscience is directed at finding *organs*; it can be assumed with Gallistel (1997, p. 86) that '…organs within the brain are neural circuits whose structure enables them to perform one particular kind of computation'. However, as we will argue, putting content into the idea that there are particular *kinds* of computation that neural circuits might be related to is a challenging question for any research programme that attempts to bridge the gap between CR and NB theories.

12. The second specialisation question asks what parts of the CR theory might be subject to an explanatory neurolinguistic account. Some parts of the CR theory might be understandable in this way, while others might not be. In fact, it could turn out to be the case that *all* properties of language are explicable in these non-NB-related terms, such that there is no 'explanatory neurolinguistics' in the sense that we intend here. We simply do not know at present.

It is useful to think of explanatory connections in terms of Marr's levels of analysis. In that system, we have at some points above collapsed the computational and algorithmic levels of analysis under the heading of CR. For present purposes, it is important to distinguish these two, because it is possible that questions of implementation – in particular, the specialisation of neural circuits – could relate the hardware in an explanatory way to the representation/algorithm, but not to the computational theory, or vice versa. Schematically, we can examine the prospects for explanatory connections for each of the pairs:

Type I: Computational/Hardware

Type II: Algorithm/Hardware

Type III: Computation/Algorithm

Of these three, Type III implicates connections between theoretical linguistics (on the assumption that that is a computational-level theory) and psycholinguistic theory (recall Section 7). It is the question of whether there are explanatory relations between particular computations found in language and the particular algorithms that execute those computations in real time. Though important, this is not our focus in this paper.

The other two types implicate links between the cognitive and neurobiological domains. Type I points to the hardware being restricted in some way, such that NB structure ***B*** computes class ***C*** of computations, and not some other class. Type II points to a scenario in which a particular NB structure is specialised for running a particular type of algorithm ***A*** (and not some other type) that is an implementation of some computation or set of computations that are part of language (see Section 9 for the example of sound localisation).

13. An immediate question that confronts Type I specialisation is what might constitute a class of computations that (different types of) NB objects could be specialised for. By this we mean that it should be asked if it makes sense to say that a particular NB structure is restricted to operate with, *e.g.* particular class of elementary operations or, *e.g.* a particular class of composed functions. (Here we are assuming a computationalist view of cognition and employing the language of computable functions, *etc.*, because we know of no other way of conceptualising this problem). This question is highly complex; for instance, it interacts with issues concerning the plasticity of neural computation, which are not, in our view, well understood at present with respect to higher cognitive function.

There are some reasons to think that relatively broad notions of specialisation, such as those that derive from automata theory might not be the correct way to go. For example, it has been argued by Gallistel and King (2009) that the computations executed by the brains of relatively 'simple' organisms (*e.g.* ants) require Turing machine computational power (this is part of a critique of 'bottom-up' views of cognition like those instantiated in connectionist research programmes). If it is the case that even the comparatively 'simple' neural circuitry found in ant brains is capable of computational operations that make it as powerful as a Turing machine, then (by definition) even simple systems are capable *in principle* of computing all computable functions. If this view is correct, then the hypotheses about specialisation in the human brain that make divisions along automata theoretic lines – for example, the hypothesis that some NB structure that is part of language could perform operations equivalent in power to a finite-state machine, but not higher – would appear to be difficult to support. If a 'simple' structure in an ant's brain can compute in a way that is equivalent to a Turing machine, is it plausible that there should be sharp computational restrictions in the neural structures that compute (subparts of) language?

The question about what computational dimensions could be used to define potential instances of specialisation is of particular interest because investigations in mathematical linguistics have made a number of explicit proposals concerning the automata theoretic power found in different parts of the CR theory of language: *e.g.* mildly context-sensitive syntax (*e.g.* Joshi, Vijay-Shankar, & Weir, 1991), or sub-regular phonology (for review see Heinz 2011a, 2011b), etc. If subparts of language are indeed restricted computationally in these (or related) ways, then we would like to know whether it is

neurobiological specialisation of the type we have introduced above that explains why these restrictions are the way they are, or something else. Or, returning to our original point, if specialisation is to be understood in classes that are not defined in automata theoretic terms, then we would like to know what theory of computation makes divisions among classes of computations that could be likely candidates for specialisation as we define it here.

Finally, it is also worth considering the possibility that Type I connections could exist for reasons that do not have to do with the NB structures that implement a computation *per se*. Computational devices do what they are programmed to do. So, it could be the case that, while the different NB structures that compute the subparts of language are *in principle* capable of computing like Turing machines, they are limited to compute only particular classes of subclasses (perhaps defined in terms of automata theory, perhaps not) for reasons that derive from constraints on the programmes that they are executing. This would be an interesting possibility, but it is one that (ultimately) leads to a further question about specialisation – namely, why certain programmes are employed in executing the procedures that make up language, and not others.

The contents of this section are highly speculative. However, we are willing to engage in this type of speculation precisely because the potential for unified explanation in Type I connections makes it the ultimate goal of looking at language in the brain.

14. Type II specialisation involves relations between Algorithms and NB. Unlike what we discussed with Type I connections in the last section, it is easier to imagine what form this might take. We might find, for instance, that an NB object ***B*** is in principle capable of computing any procedure, but it has properties that – due to its hardware configuration – make it in practice optimal for computing certain procedures, and not others. In a sense, this might be like asking whether specialisation for CR is of the type that is found with, e.g. the video cards that are part of most computers – where in principle either the CPU or the video card could compute any computable function, but where in practice the video card is optimised to execute procedures related to one particular computational subdomain.

15. The challenges that we outline in this paper are directed both at CR and at NB investigations of language. For the former, our view is that as more and more progress is made in understanding the computational properties of language, it is necessary to ask *why* those particular properties (and not others) are found there; and looking at the brain seems like a good way to find potential answers. For the latter, we see the ultimate goals of neurolinguistics in terms of unified theories of CR and NB; and we see three major ways in which CR and NB could be found to be connected when research is pursued in an integrated way.

The field has reached a point where we can see the beginning of what we have called correlational and integrated neurolinguistics. Explanatory neurolinguistics, on the other hand, remains something like a shadowy possibility on the horizon. It could be that in asking for the kinds of connections found with explanatory neurolinguistics, we have simply set the bar too high, or that we have simply guessed incorrectly that some aspects of NB will explain some parts of CR. This old world keeps spinning round, and time will tell if these speculations yield novel experimental insight.

Funding

David Embick's work is supported by [grant number NIH R01HD073258], David Poeppel's work is supported by [grant number NIH 2R01DC05660].

Notes

1. A comprehensive treatment would also take into account the substantial body of work from neuropsychology, which has a much longer history.
2. For general perspectives that are congenial to the one advanced here, see Carandini (2012) and Mausfeld (2012), as well as Henson (2005) and Page (2006). Marantz (2005) also argues for a similar perspective for neurolinguistics.
3. The idea that language can be approached in these terms is stressed in some recent work under the heading of *Biolinguistics* (see e.g. Chomsky (2005)). While we are sympathetic to many of the (mostly programmatic) suggestions in Chomsky's work, in practice much of the work that falls under that particular heading differs markedly in focus from the programme that we advance here.
4. On this theme, a difficult question is whether or not the CR theory should limit itself to hypotheses that are 'neurobiologically plausible', e.g. as was argued on occasion for connectionist architectures. We do not believe that it is advisable to do so at present, given our limited understanding of how any information is represented and computed in the brain. See Gallistel and King (2009) for extensive debate as well as section 11.
5. For a related discussion of the role of imaging data and its relation to psychological theories see Coltheart (2006).
6. For an important discussion of part of the history of this dynamic, see Phillips (1996) and Marantz (2005) on the *derivational theory of complexity* and its relation to the development of psycholinguistics and theoretical linguistics.
7. See Embick and Marantz (2005), Stockall and Marantz (2006), and Marantz (2013) for some specific proposals concerning how theories of morphology connect with psycho- and neurolinguistic data.
8. Part of what is at issue with the cognitive vs. brain data concerns what Chomsky (2000) calls *methodological dualism*.
9. Anatomically informed work is of this type can be productive. Our point is that, particularly in the domain of language, the area-based reasoning (and its concomitant assumptions about linguistic representation that implicate the GMP above) is prominent.

10. A further argument is that some approaches (e.g. 'connectionism') are restricted to mechanisms that are not computationally strong enough to compute what is required by the CR theory. See Gallistel and King for an argument of this type from the domain of memory.

References

Carandini, M. (2012). From circuits to behavior: A bridge too far? *Nature Reviews Neuroscience*, *15*, 507–509.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

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

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

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. Cambridge, MA: MIT Press.

Coltheart, M. (2006). What has functional neuroimaging told us about the Mind (so far)? *Cortex*, *42*, 323–331. doi:10.1016/S0010-9452(08)70358-7

Embick, D. (2010a). *Localism versus globalism in morphology and phonology*. Cambridge, MA: MIT Press.

Embick, D. (2010b). *Stem alternations and stem distributions* (Unpublished manuscript). University of Pennsylvania. http://www.ling.upenn.edu/~embick/stem-ms-10.pdf

Embick, D., & Marantz, A. (2005). Cognitive neuroscience and the English past tense. *Brain and Language*, *93*, 243–247. doi:10.1016/j.bandl.2004.10.003

Embick, D., & Poeppel, D. (2005). Mapping syntax using imaging: Problems and prospects for the study of neurolinguistic computation. In K. Brown (Ed.), *Encyclopedia of language and linguistics* (2nd ed.). Oxford: Elsevier.

Fiorentino, R., & Poeppel, D. (2007). Compound words and structure in the Lexicon. *Language and Cognitive Processes*, *22*, 953–1000. doi:10.1080/01690960701190215

Gallistel, C. R. (1997). Neurons and memory. In Michael S. Gazzaniga (Ed.), *Conversations in the cognitive neurosciences*. Cambridge, MA: MIT Press.

Gallistel, C. R., & King, A. P. (2009). *Memory and the computational brain: Why cognitive science will transform neuroscience*. Chichester: Wiley-Blackwell.

Grothe, B. (2003). New roles for synaptic inhibition in sound localization. *Nature Reviews Neuroscience*, *4*, 540–550. doi:10.1038/nrn1136

Halle, M. (1959). *The sound pattern of Russian mouton*. s-Gravenhage: Mouton & Company.

Heinz, J. (2011a). Computational phonology – Part I: Foundations. *Language and Linguistics Compass*, *5*, 140–152.

Heinz, J. (2011b). Computational phonology – Part II: Grammars, learning, and the future. *Language and Linguistics Compass*, *5*, 153–168. doi:10.1111/j.1749-818X.2011.00268.x

Henson, R. (2005). What can functional neuroimaging tell the experimental psychologist? *Quarterly Journal of Experimental Psychology A*, *58*, 193–233. doi:10.1080/02724980443000502

Jakobson, R., Fant, G., & Halle, M. (1951). *Preliminaries to speech analysis. The distinctive features and their correlates*. Cambridge, MA: MIT Press.

Jeffress, L. A. (1948). A place theory of sound localization. *Journal of comparative and physiological psychology*, *41*, 35–39. doi:10.1037/h0061495

Joshi, A. K., Vijay-Shanker, K., & Weir, D. (1991). The convergence of mildly context-sensitive grammar formalisms. In T. Wasow & P. Sells (Eds.), *The processing of linguistic structure* (pp. 31-81). Cambridge, MA: MIT Press.

Lewis, S., & Phillips, C. (in press). Aligning grammatical theories and language processing models. *Journal of Psycholinguistic Research*.

Marantz, A. (2005). Generative linguistics within the cognitive neuroscience of language. *The Linguistic Review*, *22*, 429–445.

Marantz, A. (2013). No escape from morphemes in morphological processing. *Language and Cognitive Processes*, *28*, 905–916. doi:10.1080/01690965.2013.779385

Marr, D. (1982). *Vision*. San Francisco, CA: WH Freeman.

Marslen-Wilson, W., & Tyler, L. K. (1998). Rules, representations, and the English past tense. *Trends in Cognitive Sciences*, *2*, 428–435. doi:10.1016/S1364-6613(98)01239-X

Marslen-Wilson, W., & Tyler, L. K. (2007). Morphology, language and the brain: The decompositional substrate for language comprehension. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *362*, 823–836.

Mausfeld, R. (2012). On some unwarranted tactic assumptions in cognitive neuroscience. *Frontiers in Psychology*, *3*, 1–13. Article 67.

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, *343*, 1006–1010.

Obleser, J., Lahiri, A., & Eulitz, C. (2004). Magnetic brain response mirrors extraction of phonological features from spoken vowels. *Journal of Cognitive Neuroscience*, *16*, 31–39.

Page, M. P. (2006). What can't functional neuroimaging tell the cognitive psychologist? *Cortex*, *42*, 428–443.

Phillips, C. (1996). *Order and structure* (PhD Thesis). Cambridge, MA: Massachusetts Institute of Technology.

Phillips, C., & Lau, E. (2004). Foundational issues. *Journal of Linguistics*, *40*, 571–591.

Phillips, C., Pellathy, T., Marantz, A., Yellin, E., Wexler, K., Poeppel, D., & Roberts, T. P. L. (2000). Auditory cortex accesses phonological categories: An MEG mismatch study. *Journal of Cognitive Neuroscience*, *12*, 1038–1055.

Poeppel, D., & Embick, D. (2005). Defining the relation between linguistics and neuroscience. In A. Cutler (Ed.), *Twenty-first century psycholinguistics: Four cornerstones* (pp. 103–120). Hillside, NJ: Lawrence Erlbaum.

Poeppel, D. (2012). The maps problem and the mapping problem: Two challenges for a cognitive neuroscience of speech and language. *Cognitive Neuropsychology*, *29*, 34–55.

Scharinger, M., Merickel, J., Riley, J., & Idsardi, W. J. (2011). Neuromagnetic evidence for a featural distinction of English consonants: Sensor-and source-space data. *Brain and language*, *116*, 71–82

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America*, *111*, 1872–1891.

Stockall, L., & Marantz, A. (2006). A single-route, full decomposition model of morphological complexity: MEG evidence. *Mental Lexicon*, *1*, 85–123.

Trubetzkoy, N. S. 1939. *Grundzüge der Phonologie* [Principles of Phonology]. Göttingen: Vandenhoek & Ruprecht.