A parallel architecture perspective on pre-activation and prediction in language processing

Falk Huettig a,b,*, Jenny Audring c, Ray Jackendoff d,e

a Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands
b Centre for Language Studies, Radboud University, Nijmegen, Netherlands
c Leiden University Centre for Linguistics, Arsenaalstraat 1, 2311 CT Leiden, Netherlands
d MIT Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA, USA
e Tufts University, Medford, MA, USA

**ABSTRACT**

Keywords: Language processing Linguistic theory Parallel architecture Prediction Psychology Psycholinguistics Representations Semantics Sentence processing Syntax

A recent trend in psycholinguistic research has been to posit prediction as an essential function of language processing. The present paper develops a linguistic perspective on viewing prediction in terms of pre-activation. We describe what predictions are and how they are produced. Our basic premises are that (a) no prediction can be made without knowledge to support it; and (b) it is therefore necessary to characterize the precise form of that knowledge, as revealed by a suitable theory of linguistic representations. We describe the Parallel Architecture (PA: Jackendoff, 2002; Jackendoff & Audring, 2020), which makes explicit our commitments about linguistic representations, and we develop an account of processing based on these representations. Crucial to our account is that what have been traditionally treated as derivational rules of grammar are formalized by the PA as lexical items, encoded in the same format as words. We then present a theory of prediction in these terms: linguistic input activates lexical items whose beginning (or incipit) corresponds to the input encountered so far; and prediction amounts to pre-activation of the as yet unheard parts of those lexical items (the remainder). Thus the generation of predictions is a natural byproduct of processing linguistic representations. We conclude that the PA perspective on pre-activation provides a plausible account of prediction in language processing that bridges linguistic and psycholinguistic theorizing.

1. What this paper is about and why

This paper brings together two theoretical topics. The first is the Parallel Architecture (PA), a linguistic theory which has been developed and defended on the basis of purely linguistic phenomena, but which also offers a relatively direct connection to psychology of language processing (Jackendoff, 1987b, 1997, 2002, 2007; Jackendoff & Audring, 2020). The second focus is an analysis of prediction in language processing – its mechanism and its function. Bringing these two foci together, we develop a general theory of how the PA bears on the understanding of prediction. Our larger goal is to bridge the gap between linguistic theory (“competence”) and psycholinguistic theory (“performance”), as a step toward an integrated theory of language and the mind/brain.

We stress from the start that our discussion is primarily on the theoretical plane. Many components of the approach are well-known in the literature; its novelty lies in large part in its synthesis of a broad range of theoretical constructs. However, from this stance we are able to offer reinterpretations of many previous experimental results and to make some suggestions for experiments to be devised.

We choose to analyze prediction because it has in recent years become an influential theoretical construct for explaining how the human mind works (e.g. Bar, 2007; Clark, 2013; Friston, 2005; Rao & Ballard, 1999; Suddendorf & Corballis, 2007). Experimental results in the cognitive (neuro)sciences are now frequently interpreted within a predictive mind framework, even if participants in experiments do not directly show any anticipation, and even though similar results have previously been interpreted very differently (cf. Ferreira & Chantavarn, 2018). In what might be called the “predictive turn”, many in the cognitive and brain sciences have come to consider the human mind a “predictive engine” or “prediction machine” (Clark, 2013; Friston, 2005). Some (e.g. Clark, 2013) have gone so far as to describe the last...
decade as a true paradigm shift in the sense of Kuhn (1970), comparable to the change from behaviorism to cognitivism in the 1950s.

In line with this general theoretical shift, psycholinguistic research has begun to posit prediction as an (or even the) essential characteristic of language processing (e.g. Altmann & Mirkovic, 2009; Dell & Chang, 2014; Falandary, Nguyen, & Spivey, 2021; Federmeier, 2007; Ferreira & Chantavarin, 2018; Ferreira & Qiu, 2021; Gibson, Bergen, & Plantadosi, 2013; Hale, 2001; Hickok, 2012; Huettig, 2015; Kuperberg & Jaeger, 2016; Levy, 2008; Norris, McQueen, & Cutler, 2016; Pickering & Gambi, 2018; Pickering & Garrod, 2013; Van Petten & Luka, 2012). In a sense, psycholinguistic theory is just catching up with research on vision, which long ago posited an important role for prediction in object recognition (Helmholz, 1856), and with research on music, where major theories have been built around the notion of expectation (Huron, 2006; Meyer, 1953; Narmour, 1977). Given the influence of this approach in the literature, it is of interest to sharpen the notion of prediction. This paper attempts to do so, using the theoretical tools of the Parallel Architecture. We put off comparison to other approaches to section 6, after we have developed our own treatment.

To be clear about what we mean here by prediction in language processing: we wish to think of it as the pre-activation of linguistic representations, before incoming bottom-up input has had a chance to activate them. This way of thinking about prediction is simple and straightforward and corresponds closely to the ordinary language sense that prediction is about what will happen in the future. We will return to this issue of definition in section 5.

Three basic questions for a theory of prediction are:

• What is the form of a prediction?
• How are predictions generated?
• What is the role of predictions in the course of language comprehension?

We find that the literature tends to emphasize the third of these. Here, however, we will focus on the first two.

Anticipating our answers to these questions in sections 3–6, we briefly lay out our approach here. We begin with some basic assumptions. First, it is generally agreed that a lexical item in long-term memory becomes activated in response to an input that it matches. Second, it is generally agreed that processing is opportunistic: activation spreads from an activated item to related [similar] items in memory. Third, it is generally agreed that lexical access is “promiscuous”: all items consistent with the current input are activated, resulting in multiple items in competition for “what is being heard.”

We propose that a prediction in language comprehension takes the form of a piece of linguistic structure in memory that has been activated prior to its possible appearance in the input, i.e. it is “pre-activated.”1 Once pre-activated, it interacts with subsequent input and with other predictions, facilitating or interfering with processing, as the case may be.

There are basically two ways to pre-activate lexical material (i.e. generate predictions). The first is through spreading activation in semantic networks (Anderson, 1983; Collins & Loftus, 1975; Hutchison, 2003) or through overlapping semantic features between concepts (McRae, de Sa, & Seidenberg, 1997).2 This case is generally termed “semantic” or “associative priming,” and its existence is not controversial. We’ll call this mechanism between-item pre-activation; we take this case up in section 4.

A second mechanism for pre-activation, which we stress in the present paper, arises from the opportunistic nature of processing: a lexical item may be activated even before it has appeared in the input in its entirety. Let us call the part of the item that has already matched the input the “incipit,”3 and the part that has not yet been encountered in the input the “remainder.” In effect, the pre-activated remainder constitutes a prediction of what is to come in the input. We’ll call this mechanism within-item pre-activation; we discuss it in section 5.

The PA-based account allows prediction to be seen as a natural consequence of the machinery involved in lexical access. We point out at the outset that the pre-activation account can be seen as complementing contemporary probabilistic approaches to predictive language processing, but from a somewhat different perspective (something we discuss in section 6). The role of the Parallel Architecture is to make the notion of prediction more precise and to emphasize its generality. In particular, the PA’s notion of an “extended lexicon” (see section 2) allows for pre-activation (hence prediction) in syntax, semantics, and phonology, and at all scales, from morphological affixation to words, syntactic constructions, and even poetry.

2. A theory of representations

2.1. The Parallel Architecture

Before discussing our approach to processing, we briefly set out the relevant tenets of the Parallel Architecture and its progeny, Simpler Syntax (Culicover & Jackendoff, 2005) and Relational Morphology (Jackendoff & Audring, 2020).

At its foundation, the PA couches the fundamental questions of linguistic theory in psycholinguistic terms:

• What linguistic entities are stored in long-term memory, and crucially, in what form are they stored? (a.k.a. “knowledge of language,” “theory of competence,” theory of representations)
• How are these entities put to use in language comprehension and production? (theory of “performance,” theory of processing)
• How are these entities acquired? (theory of acquisition)
• What is the innate foundation that supports language acquisition? (“Universal Grammar”)
• How does language fit into the rest of the mind?

This section addresses the first two and the last of these.

The feature that gives the framework its name is its treatment of phonology, syntax, and semantics as independent combinatorial systems, operating in parallel. Each of these levels of representation has its own generative capacity, built out of its own characteristic units. Fig. 1 sketches the overall architecture. The double-headed arrows represent interface links: correspondences between levels rather than derivations from one level to another. A well-formed sentence has well-formed structures at each level, plus well-formed links among the structures.

For the simplest sort of example, the word cat consists of a piece of semantic structure (the meaning of the word), a piece of phonological structure (/kæt/), and the syntactic category Noun – plus interface links

1 A similar characterization appears in Chow, Smith, Lau, & Phillips, 2015: “…we use ‘prediction’ to refer to pre-activation of stored representations before the bottom-up input arises.” See section 5 for other terminology.
2 We are agnostic as to whether semantic priming is due to associative strength, feature overlap, or a combination of both.
3 We borrow the term incipit from musical analysis, where it denotes the opening motive of a melody, where it denotes the opening motive of a melody, where it denotes the opening motive of a melody.
4 Jackendoff, 1997, 2002 uses the term “correspondence rules” in the same sense.
between them. These links are noted by co-subscripting, as in (1); one can think of the subscripts as marking the ends of association lines.

(1) Semantics: [CAT]
   Syntax: \( N_i \)
   Phonology: /kæt/

Thus a word consists of a triple of representations, linked across the three levels. Most linguistic theories share this three-level conception of a word; they vary primarily in how they conceptualize the links between the three.²

This account has direct implications for processing. In language comprehension, an input /kæt/ activates the phonological layer of (1). Activation then spreads along the interface link to syntactic structure \( N_i \) and to the meaning CAT (not necessarily in that order), and they remain linked as the rest of the utterance is processed. If we think of semantics as “high-level” and phonology as “low-level,” then comprehension is fundamentally “bottom up” (without precluding “top-down” feedback). In language production, the flow of activation goes in the other direction: it begins with the intended message CAT, and it spreads along the interface link to activate the associated noun pronounced /kæt/. Thus we can think of production as fundamentally “top-down” (without precluding “bottom-up” feedback).³

The PA’s treatment of the relation among levels of representation differs from the widespread premise in generative linguistics that recursive syntax is the primary source of combinatoriality in linguistic structure, and that semantics and phonology are derived from it (Berkvick & Chomsky, 2016; Chomsky, 1965, 1995).⁴ Although semantics and phonology are unquestionably related to syntax, they each have independent properties that cannot be derived from syntax (Jackendoff, 1997, 2002). Moreover, there exist direct links between semantics and phonology, independent of syntax, for instance the relation between focus in semantics and intonation in phonology. Moreover, for the Parallel Architecture, the use of recursion to achieve the “infinite use of finite means” pertains not only to syntax but to semantics as well – the content of the messages that syntax and phonology convey.

2.2. The extended lexicon

Along with constraint-based theories such as Construction Grammar (Croft, 2001; Goldberg, 1995) and Head-Driven Phrase Structure Grammar (Pollard & Sag, 1994), the Parallel Architecture differs from classical generative linguistics in regarding the lexicon as far more than the traditional repository of words. For one thing, one’s knowledge of language has to encompass idioms – pieces of phonological and syntactic structure that contain multiple words but with a meaning that cannot be constructed from the meanings of their parts. For example, (2) illustrates the PA’s notation for the lexical entry of our dear old friend kick the bucket. The idiom’s three levels are tied together into a lexical item by subscript 2. Within this item, subscripts 3, 4, and 5 link the syntax to the phonology of the individual words – but they have no link to the semantics.

(2) Semantics: [DIE]:
   Syntax: \( \forall V_{1} \left[ \forall D_{0} \exists N_{1} \right] \);
   Phonology: \( \text{kick} \; \delta_{\text{D}} \; \text{kick} /_{2} \)

In addition to thousands of idioms, the lexicon must contain vast numbers of other multiword items: clichés such as light as a feather and red as a beet, frequent fixed expressions such as I think so, and collocations that have literal interpretations but are recognized as the “right way” to say things, e.g. phrases like black and white rather than #white and black (Christiansen & Arnon, 2017; Jackendoff, 1997; Pawley & Syder, 1983; Wray, 2002). These are all facts of English that one has to learn and store.

The lexicon also includes syntactic constructions that are linked to idiosyncratic meanings. An example is the N of an N pattern illustrated in (3) (Booij, 2002):

(3) a travesty of an experiment (= ‘an experiment that was a travesty’)
   that gem of a result (= ‘that result, which was a gem’)

In the canonical semantics for this syntactic structure, the syntactic head is also the semantic head. For instance, a picture of a cat denotes a picture, not a cat. But in (3), the syntactic heads travesty and gem are understood as semantic modifiers; they offer a negative or positive evaluation of the semantic heads experiment and result, which are expressed as syntactic dependents.⁵

Idiomatic patterns like (3) must be learned and listed in the lexicon; they are the stock in trade of Construction Grammar (Croft, 2001; Fillmore, 1988; Goldberg, 1995; Hoffmann & Trousdale, 2013; Jackendoff, 1990). Thus all of these heterogeneous types of lexical entry – words, collocations, idioms, and meaningful syntactic constructions – consist of pieces of semantic, syntactic, and phonological structure, bound together by interface links.

Another crucial difference between the Parallel Architecture and traditional generative linguistics lies in the status of rules of grammar. For instance, consider the regular plural in English. In traditional generative grammar, the formation of plurals is governed by a derivational rule roughly of the form “To form the plural of a noun, add -s.” The counterpart in the PA is a schema of the form (4).

(4) Semantics: [PLUR \((X_{i})\)]
   Morphosyntax: \[N_{1}; \text{PLUR}_{1} \]
   Phonology: \( \ldots \times \it{\text{bs}} /_{y} \)

Like the entries in (1) for cat and in (2) for kick the bucket, schema (4) consists of a piece of semantics, a piece of (morpho-)syntax, and a piece of phonology, linked by subscripts. The difference is that parts of its structure are variables: (4) says that a multiplicity (PLUR) of any sort of entity (X) can be expressed by a noun (N) plus a plural affix (PLUR), the combination being pronounced in whatever way the noun is pronounced, followed by the phoneme /s/. The plural form cats is produced by instantiating the variables in (4) – including the variable subscripts x and y – with the corresponding pieces of (1), resulting in the structure (5).

(5) Semantics: [PLUR \((\text{CAT})\)]
   Morphosyntax: \[N_{1}; \text{PLUR}_{1} \]
   Phonology: \( /_{x} \text{kat} /_{y} \)

Moreover, (4) can also be instantiated with newly encountered nouns, to spontaneously produce novel expressions such as wugs and

---

² It should be noted that homonyms have separate lexical entries. Hence one should not speak of “the ambiguous word bank,” as if there is a single word. Rather, there are two lexical entries that happen to have the same syntax and phonology but differ in meaning.

³ Another use of the terms “top-down” and “bottom-up” pertains to relative height in a syntactic tree structure, i.e. within a single level of representation. The two uses must not be confused. See section 5.3.

⁴ In earlier generative grammar, the generative power of language was invested in syntactic phrase structure rules and syntactic transformations. In more recent theory (e.g. Berkvick & Chomsky, 2016 and Chomsky, 1995), the generative capacity comes from the syntactic operation Merge.
coelacanths. The variables are what make it possible to create composite structures on the semantic, syntactic, and phonological levels. Unlike in traditional generative grammar, syntactic composition has no priority over the other two.

More generally, all rules of grammar can be restated in schema form: they are essentially in the same format as words, except that some of their structure is made up of variables. This approach extends even to syntactic phrase structure rules, such as that for the English noun phrase, approximated in (6). This is a piece of linguistic structure that involves only one level of structure and that consists entirely of variables. One can think of it as a “treetet” in the sense of Fodor, 1998 and Tree-Adjoining Grammar (Joshi, 1987).

(6) Syntax: \[ \text{NP Det } \langle A \rangle \text{ N } \ldots \]

In essence, then, there need be no further distinction between words and rules of grammar: they belong together in a single system that might be called the “extended lexicon” (Construction Grammar often uses the term “construction” in the same sense). In addition to Jackendoff, 2002, Culicover & Jackendoff, 2005, and Jackendoff & Audring, 2020, this point is also argued on linguistic grounds by e.g. Fillmore, 1988, Langacker, 1987, and Croft, 2001, and on psycholinguistic grounds by Bates & Goodman, 1997.

Schemas fulfill the traditional function of rules – that of creating an unlimited number of novel structures – through the operation of unification (Shieber, 1986). Unification instantiates a schema’s variables with further material, as seen in the composition of cats above. The result of unifying a word and a schema is a composite that shares all the features they have in common and preserves all the features that are distinct. We call this use the generative function of schemas. Hence the composition of a sentence involves “clipping together” an ensemble of stored pieces through unification; this provides a straightforward implementation of productivity in terms of the instantiation of variables in stored structures.

The PA lexicon as envisioned by Jackendoff & Audring, 2020 differs from the traditional lexicon in yet another respect: it encodes relations among lexical items explicitly. For example, consider a pair of words like laugh and laughter. The string -ter looks like a suffix, but it occurs only attached to the word laugh. It would be peculiar to posit a traditional rule along the lines of “to form a noun based on laugh, add ter”; a rule that applies only to a single item is no rule at all. In contrast, the PA treatment relates laugh and laughter as in (7). Note in particular that laughter is stored with internal structure on all three levels.

(7) Semantics: a. [LAUGHs]
   Morphosyntax: V
   Phonology: /laɪf/<

Here, subscript 8 links the three levels of laugh, and similarly, subscript 9 links the three levels of laughter. But subscript 8 also links laugh to the base of laughter, marking the two as the same. We call this connection a relational link. It is used not to derive laughter, but rather to explicitly encode the relation between the two lexical items, again on all three levels. The presence of this relation “supports” or “motivates” laughter: it makes it less arbitrary than a word like hurricane that lacks internal structure. Laughter is easier to learn, then, because it has a previously known part; and it is easier to process, because of the extra activation that comes from laugh (see section 3).

Schemas too can participate in relational links. Consider the idiomatic expressions in (8), which all contain the plural -s suffix.

(8) raining cats and dogs, holding hands, odds and ends, best regards

The full meanings of these expressions cannot be built up from the meanings of their parts, so the expressions must be learned and stored. But that does not entail that they are stored as holistic unstructured units. In particular, the plural nouns are still standard plural nouns, even though they are not spontaneously generated. This generalization can be captured by establishing relational links between the plural schema (4) and these idiomatic stored plurals. Again, the intuition is that the relational link to the schema makes these idioms easier to learn and process.

There is an important consequence: schemas are used not only to generate novel structures, but also to support items that are stored. We call this the relational function of schemas. In traditional rule-based approaches, rules play only a generative role, while the relational role, if even mentioned, is fulfilled by “lexical redundancy rules,” “analogy,” or less structured associations (e.g. in Pinker, 1999). The PA claims that these two functions can both be performed by a single schema, as we have seen with the plural.

Furthermore, many schemas can be used only in the relational role – that is, all of their instances have to be listed. Such schemas are responsible for nonproductive patterns such as the English sing/sang alternation, which has about a dozen instances. As a result, there is no essential difference in form between schemas for productive and nonproductive patterns – only whether they can be used generatively or not. This yields a novel perspective on the problem of productivity, which of course bears on the infamous “past tense debate” (e.g. Bybee & Mester, 1983; Rumelhart & McClelland, 1986; Pinker & Prince, 1988; Plunkett & Marchman, 1991; Clahsen, 1999; Pinker, 1999; see Jackendoff & Audring, 2020 for more detail on the present solution).

To sum up, the extended lexicon is a single system that stores not only words, but also idioms, collocations, and schemas. Schemas are stated in the same terms as words, namely as pieces of linguistic structure – semantics, (morpho)syntax, and phonology – connected by interface links where appropriate. They differ from words in that they have variables that must be instantiated in constructing an utterance. Connections among words and schemas are encoded as relational links, which mark parts of related items as the same.

While the Parallel Architecture deviates considerably from traditional generative grammar, its approach to grammar and stored representations is very much in tune with contemporary usage-based approaches such as Tomasello, 2003, and especially with Construction Grammar (Boas & Sag, 2012; Croft, 2001; Goldberg, 1995; Hoffmann & Trousdale, 2013) and Construction Morphology (Bouij, 2010). One great advantage of this approach is that words and schemas are stored in a common format that permits unification.

Important to the Parallel Architecture is that it is not simply

---

9 The free application of (4) is blocked by stored nonproductive plurals such as children and vertices. For detailed treatment of a parallel case, the nonproductive past tense forms in English such as sang and brought, see Jackendoff & Audring, 2020, especially chapter 5.

10 If it seems suspicious for a schema potentially to involve only one level of representation, note that phonotactic constraints similarly amount to schemas that involve only phonological structure.

11 And perhaps in laughter, but less transparently, because the stem *slaugh-* does not exist.

12 Here the connection is not between shared subscripts, but rather between the variable subscripts in the schema (x and y in (4)) and the constant subscripts in its instances, which are different for every instance. For example, in the plural form cats in (5), the constant subscripts are 1 and 7, in place of the variables in (4).

13 The subcategorization and selectional features found in many words are in effect variables too (see example (11) in section 5.3), further undermining the difference between words and rules.

14 In Construction Grammar and HPSG, lexical items are related through inheritance, by which one item can be taken as a special instance of another. Relational links are broader in their application than inheritance, in that they can also relate items that are similar, yet neither is a special case of the other. See Jackendoff & Audring, 2020, especially chapter 3.
proposing a notation for describing languages. The claim is that these linguistic structures (or the neural equivalents thereof, which are at present unknown) are instantiated in the brain. These structures include not only phonological, (morpho)syntactic, and semantic/conceptual structure, but also the interface links and relational links that tie lexical items together. Thus the extended lexicon – in effect knowledge of language – is taken to be a richly interconnected network in the brain.

### 2.3. Beyond language

A brief digression: An important reason to favor the Parallel Architecture approach to language is that it offers a natural account of the relation of language to other cognitive capacities. Consider the fact that we can talk about what we see. This requires the existence of an informational conduit between linguistic meaning and those higher-level visual representations that encode one’s understanding of the visually available aspects of the world. For instance, to understand the sentence *That is a cat*, accompanied by a pointing gesture, one must establish a connection between the word *cat* (as in example (1)), one’s knowledge of what cats look like, and one’s representation of something in the visual field that can serve as the referent of the demonstrative that.

Such a connection cannot be established by deriving visual representations from linguistic representations (in particular from syntax) – or vice versa. The PA instead posits that this connection is encoded in terms of a further layer of interface links that specify the respects in which linguistic meaning corresponds to visual understanding – a straightforward extension of the formalism (Jackendoff, 1987a, 1987b, 2002). This is consistent with a great deal of evidence from the visual world paradigm in psycholinguistics (Cooper, 1974; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995), which has revealed close interaction between language and vision as well as rapid activation of visual representations from linguistic input (Huetting & Altmann, 2007; see Huetting, Rommers, & Meyer, 2011, for review).

This approach can be further extended by observing that high-level visual representations and language also interact with haptic representations, which likewise encode shape and spatial layout (Fowler & Dekle, 1991; Gick, Jöhnnsdóttir, Gibraiel, & Mühlbauer, 2008; Sato, Cavé, Ménard, & Brasseur, 2010). They must also interact with proprioseption, which encodes the spatial configuration of one’s own body, and which is therefore crucial in planning and executing action (Lackner & Dizio, 2000). Fig. 2 sketches all these connections. At the top are the components involved in language, and below that are forms of representation involved in perception and nonverbal cognition.

Again there is no question of deriving any of these representations from the others; their interaction ought to be construed in terms of interface links. Thus the PA formalism offers the possibility that the same basic mechanism that connects the representational levels of language to each other also connects language to other cognitive domains, and it connects other cognitive domains to each other as well (cf. Prinz, 1990).

Moreover, one could delete the linguistic components of Fig. 2 and get a realistic architecture for a nonhuman primate (though of course more limited in the conceptual department). Monkeys and apes negotiate their spatial environment using visual, haptic, and proprioseptive input, and they negotiate their social environment with surprising conceptual sophistication (Cheney & Seyfarth, 1996; Köhler, 1927). We might add that bats presumably have an extra pathway to spatial structure via echolocation.

We take these considerations to be an important big-picture advantage of the Parallel Architecture over other proposed architectures for the language capacity, particularly those in which syntactic structure is the sole source of combinatoriality.

### 3. Standard assumptions about processing

Moving toward an account of prediction, we now embed the PA theory of linguistic representations in a partial theory of language processing, based in part on Jackendoff, 2002, 2007; Jackendoff & Audring, 2020. The gist of the argument is that familiar constructs in theories of processing can be readily interpreted in terms of the PA’s representational theory, and that this interpretation helps sharpen our understanding of processing. Hence the Parallel Architecture leads to a graceful connection between “competence” and “performance,” rather than the firewall between them advocated by Chomsky, 1965 and maintained within mainstream generative grammar.

Section 1 listed some standard tenets of psycholinguistic theory. Here we elaborate them in the context of the Parallel Architecture. (We acknowledge that they are not universally accepted.)

An important tenet is that language comprehension involves two components: lexical access and integration. In the first of these, incoming phonological input activates identical (or sufficiently similar) pieces of phonological structure in the lexicon (Allopenna, Magnuson, & Tanenhaus, 1998; Marslen-Wilson, 1987). These pieces pass activation to corresponding structures in the lexicon – to both syntactic structure (Cleland & Pickering, 2003; Pickering & Branigan, 1998) and semantic structure (Huetting & Altmann, 2005; Meyer & Schwaneveldt, 1971; Yee & Sedivy, 2006). From the PA perspective, this transmission of activation takes place specifically via the interface links stored in the lexicon. In the second component of comprehension, the processor attempts to integrate the accessed structures with the current hypothesis or hypotheses about the syntax and semantics that have been built on the basis of previous input. Here we will be concerned only with the lexical access component of processing – i.e. the activation of candidates for “what is heard.” (For some discussion of integration in the PA framework, see Jackendoff, 2002, 2007; Jackendoff & Audring, 2020.)

A basic feature of lexical activation is the well-known role of frequency: more frequent items activate more quickly and/or more strongly than less frequent items. In concurrence with much of the

---

15 We take the nodes of the lexical network to be internally structured and contentful, according to linguistic principles, as seen in (1), (2), (4), (5), and (7). One can consider entire lexical items to be nodes, or individual levels of lexical items, or particular units within an individual level. Much of the modeling literature, especially work on neural networks, assumes atomic nodes whose only content is their activation level and their connections to other nodes. We reject this assumption, on grounds detailed in Marcus, 1998, 2001, and Jackendoff, 2002 (section 3.5) – basically, because such a network allows no account of compositionality. How our network, or any other proposed network, is implemented by neurons is unknown.

16 A caution, partly recapitulating note 15: This account makes no claims about neural implementation, e.g. that interface links are implemented by axons, or that activation spreads along axons, or that nodes in the lexical network are implemented in single neurons or in assemblies of neurons. The difficult problem of neural implementation is common to all current theories of language storage and processing, and we don’t aspire to solve it here.
psycholinguistic and neuroscientific literature (e.g. Collins & Loftus, 1975; Grainger & Segui, 1990; Luce, Goldinger, Auer, & Vinevitch, 2000; Pierrehumbert, 2001), we take the frequency of an item in a corpus to be a proxy for its “resting activation” or its “lexical strength” in the brain – its readiness to respond to incoming activation. Under the usual assumption that any use of a word augments its resting activation, an item that is encountered more frequently will have a higher resting activation. Therefore it will have a livelier response to subsequent activation, and it will be able to (somewhat stochastically) outcompete other candidate items for “what is being heard” (for experimental evidence, see e.g. Dahan, Magnuson, & Tanenhaus, 2001; Marslen-Wilson, 1987).

In concurrence with much of the literature, we assume that the course of processing is opportunistic or incremental, in the sense that phonological, syntactic, and semantic information is brought to bear whenever it becomes available (e.g. Altmann & Steedman, 1988; Marslen-Wilson, 1975; Tanenhaus et al., 1995). Moreover, consistent with a wealth of evidence from the “visual world” paradigm, even visual information can be brought to bear on syntactic parsing, if available in time (Tanenhaus et al., 1995; Huettig et al., 2011 for review). From the PA perspective, this amounts again to passing activation through interface links, this time from visual/spatial levels of representation to semantic structure.

It is important to note that passing activation through interface links still does take time, which affects the overall time-course of processing. For example, in comprehension, a word cannot spread its activation to semantic associates until its own semantics has been activated by its phonology, via its interface links. (See the treatment of chair priming sit below.)

We assume further (though this is still somewhat controversial (Ferreira & Patson, 2007)) that processing is promiscuous, i.e. multiple possibilities for lexical identification, syntactic parsing, and semantic interpretation are processed in parallel. These possibilities, weighted by resting activation, similarity to the input, and priming (including priming by context), compete with each other for “what is being heard.” This mode of processing pertains to all levels of structure, including for instance “cohort” processes in phonology (Marslen-Wilson, 1987) as well as syntactic parsing (MacDonald, Pearlmutter, & Seidenberg, 1994; Swinney, 1979; Tanenhaus, Leiman, & Seidenberg, 1979; Trueswell, Tanenhaus, & Kello, 1993). We return to priming and cohort processes below.

In addition to these widely accepted features of language comprehension, we briefly suggest four potential refinements that are useful to this account (Jackendoff & Aurding, 2020) and possibly to others. First, we assume that an item being heard for the first time is stored with some initial resting activation, whose strength may be modulated by factors such as newsworthiness and fashion.

Second, augmentation of an item’s resting activation had better not continue without limit, even for extremely frequent items. In line with accounts of asymptotic learning, we propose that augmentation of an item raises its resting activation asymptotically toward a ceiling. Highly frequent items are boosted by further exposure only imperceptibly, if at all; whereas items that have been only seldom encountered, and hence have a much lower resting activation, receive a more substantial boost (cf. Morton, 1969; Seidenberg & McClelland, 1990; Van Orden, Pennington, & Stone, 1990). Such augmentation of resting activation by asymptotic learning is compatible with accounts that suggest that increasing practice eventually results in progressively smaller effects, which can be represented either as an exponential function (Heathcote, Brown, & Mewhort, 2000; Myung, Kim, & Pitt, 2000) or a power function (Anderson, 1982; Logan, 1990; Newell & Rosenbloom, 1981).

Third, we assume that resting activation slowly decrements with disuse; if it falls below some threshold, the lexical item is effectively forgotten (Holstädter & Mitchell, 1995; Kaptinski, 2007).

Fourth, we propose that the brain is inherently a noisy computational environment, so that “resting activation” really represents a stochastic distribution of activity over time. This proposition seems to be taken for granted in research at the neural level (e.g. Dinse, Heeger, & Behrmann, 2015). The presence of noise is the reason why any psychological experiment needs substantial numbers of subjects and stimuli to achieve statistical reliability; it is also the reason why even the very best basketball players cannot sink a free throw every time; in language production it is a source of speech errors. In the present context, decrementing the resting activation of a disused lexical item eventually results in its being obscured by the noise. There is no need to posit a strict threshold.

The assumption of promiscuity raises the threat of computational explosion from too many possibilities being considered at once, in parallel. Bayesian approaches to cognition and language face a similar problem: it would seem beyond the capabilities of even something as complex as the brain to make exact calculations that consider the vast number of possible probabilities (Kwisthout, wareham, & van Rooij, 2011). Bayesian approaches overcome this problem by assuming that explicit consideration of a whole probability distribution is not required, and that sampling from such a distribution is a good approximation to the laws of probability (Sanborn & Chater, 2016). The present approach avoids computational intractability by postulating some threshold of activation below which potential hypotheses cannot achieve candidate status. We conjecture that the threshold in question is a consequence of the noisy character of neural computation: items can only achieve candidacy if their activation is sufficient to stand out from the noise.

4. Between-item pre-activation and identity priming

Another standard assumption in the psycholinguistic literature is that activation of a lexical item spreads to similar or related items, due either to learned association in semantic networks (Anderson, 1983; Collins & Loftus, 1975; Hutchison, 2003) or overlapping semantic features between concepts (McRae et al., 1997). Activation by this route is generally referred to as semantic priming; it constitutes one of the two types of pre-activation mentioned in section 1.

The Parallel Architecture allows us to be more precise, in that it has a specific hook for identifying the loci and magnitude of spreading activation: activity spreads through relational links between items – on any or all of three levels of representation. The intensity of activity that spreads from one item to another is determined not only by the level of activation of the “donor” item, but also by the degree to which the items in question are linked relationally. It therefore follows that more activation will be spread between items whose phonological relation is relatively transparent, such as joy/joyous, compared to a less closely related pair such as malice/malicious, whose phonological relation is more tenuous, thanks to the differences in stress and vowel quality. This conclusion is borne out experimentally (Pinker, 1999; Zwiterlood, 2018). Similarly, on the semantic plane, king spreads more activation to the closely related queen than to, say, mailman, with which it shares far fewer semantic features and no phonological features.

Activation can also spread between items that are related only morphologically (i.e. phonologically and syntactically, but not semantically), such as rectis/recit, expose/exposition, and gorgeous/gorge (Amenta & Grepaldi, 2012; Diependaele, Dinabertia, Morris, & Keulers, 2011; Giraud & Grainger, 2003). Activation can even spread between items that are related only phonologically, such as corner/corn and brother/broth – though this is possible only through spurious identification of -er with the homonymous affix, and the activation is faint...
Semantics: CHAIR → SIT
Phonology: /ʃɛr/ → /sɪt/
Input: /ʃɛr/

Fig. 3. Chair primes sit. Vertical arrows denote activation that spreads via interface links. The horizontal arrow denotes activation that spreads via a relational link in semantics.

enough to require extreme measures to detect it (Rastle, Davis, & New, 2004). In addition, there is some evidence that words spread purely phonological activation to words that rhyme with them (Allopenna et al., 1998).

Activation can furthermore spread on the basis of associative or semantic relations alone. For instance, part of the meaning of the noun chair is that its function is for sitting. It is therefore expected to prime and be primed by the verb sit, despite the absence of phonological and syntactic similarity. It is not clear whether the notion of sitting is part of the lexical entry for chair or just part of one’s “world knowledge” about chairs. But in either case, this aspect of chairs is connected to the semantics of the lexical entry of sit via a relational link. Fig. 3 illustrates the flow of activation in a situation where hearing chair in the input primes sit. (We show only the semantic and phonological levels.)

The input is phonological, activating the phonology /ʃɛr/ in the lexicon. This passes activation through an interface link to the semantics CHAIR (what might be called “bottom-up” activation). From here, activation spreads “horizontally” through its relational link to SIT. SIT in turn passes activation “top-down” through its interface link to the phonology /sɪt/, making it more susceptible to activation from subsequent input. Notice that the only path available for activation to spread is the one in Fig. 3: /ʃɛr/ can’t directly activate either SIT or /sɪt/. (We will see more complex interactions in section 5.)

This account extends beyond such word-to-word semantic priming to a broader sense of priming which includes discourse and even nonlinguistic understanding of the current situation, in line with psycholinguistic evidence (e.g. Nieuwland & Van Berkum, 2006). In particular, discourse understanding can involve prediction in visual/spatial structure as well as linguistic structure. For example, in the situation depicted in Fig. 4, the exclamation Don’t might well activate the word jump. We do not explicitly formalize this here, but we do assume that discourse, event and world knowledge, and visual and spatial information can prime lexical items, contingent on the contextual situation.

Now recall that the Parallel Architecture’s extended lexicon stores idioms, collocations and schemas right alongside of words. These are all pieces of linguistic structure — stored knowledge — and they all involve both interface links, which connect their levels, and relational links, which connect them to other stored items. The consequence for processing is that all principles of lexical activation and lexical access should apply to these items in the same way as they apply to words.

Note that this is not a feature of traditional accounts in which the lexicon and the grammar are taken to be distinct. For instance, while the problem of lexical access is taken to be central, the literature does not typically recognize the parallel notion of “rule access”, i.e. choosing what rule to apply in a derivation. Rather, standard accounts are stated in terms of choosing among structures — the outputs of rule application, such as high vs. low PP attachment in The woman saw the man with a telescope. This is typically treated as quite a different process from accessing words. In the PA framework by contrast, the construction or parsing of a sentence involves activating and selecting “treelets” such as the noun phrase schema (6), through the very same process that activates and selects words. Thus choosing among structures is altogether natural. Structural ambiguities should function just like lexical ambiguities.18 Indeed, most current frameworks in psycholinguistics appear to acknowledge there is no hard distinction between rules and memory-based access (in line with constraint-satisfaction accounts of sentence processing, MacDonald et al., 1994). By putting the conception of resting activation together with the status of schemas as lexical items, it follows that more frequent syntactic schemas (e.g. more frequent syntactic constructions) have a higher resting activation, making their response more robust in both comprehension and production; this is in line with evidence in the psycholinguistic literature (e.g. Juliano & Tanenhaus, 1993; MacDonald et al., 1994; Mitchell, Cueto, Corley, & Brysbaert, 1995).

Given the status of schemas, activation spreads not only between one word and another, but also between a word and the schema(s) that it is an instance of. This bears on the long-standing disputes about the processing of multimorphemic words (e.g. Taft, 2004; Butterworth, 1983; Clahsen, 1999; Baayen, Dijkstra, & Schreuder, 1997; for surveys, see Zwitserlood, 2018 and Gagné & Spalding, 2009). For instance, the PA claims that the word widen is stored — with its internal structure (Jackendoff & Audring, 2020). Activating it spreads activation to both the word wide and the schema that supports the pattern widen/blacken/harden/tighten/etc.19 These activations reinforce that of widen itself, increasing the processor’s commitment to this as “the word being heard.” Hence the judgment is faster and/or more robust.20

This analysis predicts that all else being equal, a stored but fully

18 The PA conception of processing as assembly of stored pieces of structure might be thought of as a marriage of “constraint-based” parsing in the sense of MacDonald et al., 1994 and (Lexicalized) Tree-Adjoining Grammar (TAG: Joshi, 1987; and LTAG: Joshi & Schabes, 1997). The former claims that all the material being assembled comes from the lexicon, with which we concur. But it builds higher-level phrase structure into the entries of individual words. For example, it claims that every noun, such as dog, is listed not as [N (dog)], but as [NP <Det> <AP> [N (dog)]], making the lexicon massively redundant. (See Jackendoff, 2007 for more discussion.) TAG, on the other hand, derives sentences from fragments of phrase structure like (6), with which we concur. But treelets are treated as formally distinct from words. In the present approach, fragments of phrase structure are stored as schemas, as in TAG; but they are in the lexicon, as in constraint-based parsing.

19 Note this schema is not productive; one is not free to make up new instances such as *crispen and *louden, the way one can with plurals such as coelacanths.

20 In probabilistic terms, the independent activity of the parts increases the probability that the word being heard is widen. See Jackendoff & Audring, 2020, for a more detailed analysis of the interaction between stored items and computation of their parts.
In addition, there can be morphological neighborhood priming: a word containing a particular affix may prime other words with the same affix directly as well as via the affix schema.
lie on a continuum and could potentially be accounted for by a single mechanism (such as simple recurrent networks, cf. Elman, 1990). Given this terminological tangle, we will retain the term “prediction” for both “between-item pre-activation” and “within-item pre-activation” and do our best to keep clear what we have in mind (see also Kuperberg and Jaeger, 2016, who consider the various terminologies, and settle on a definition not unlike ours).

5.2. Two examples

We illustrate within-item pre-activation with two examples. The first is a typical cohort process in phonology (Marslen-Wilson, 1987; Switzerland, 1989), illustrated in Fig. 5 (next page).

Here is what happens:

- An input /kæp .../ activates the phonological unit /kæp/ in the lexicon, which activates the semantics CAP “bottom up” via an interface link. CAP then spreads activation “horizontally” to its semantic associate HAT, via a semantic relational link – which then activates the phonology /hæt/ “top-down”. This is exactly like chair pre-activating sit in Fig. 3 above.
- However, since activation is promiscuous, the input /kæp .../ also activates the phonology of other lexical entries such as captain, cappuccino, capitalism, capstan, and so on, for which /kæp/ functions as an incipit.
- Now: even though the phonological remainders /-tən/, /-ətʃnəʊ/, and so on have not been heard, they are pre-activated, and the completed phonology of the words activates their semantics, via the up arrows in Fig. 5. These in turn spread activation to their semantic associates, via the horizontal arrows.
- Finally, the semantic associates activate their own phonology, via the down arrows in Fig. 5. Thus the pre-activated phonological fragments /-tən/ and /-ətʃnəʊ/, and the semantics CAPTAIN and CAPPUCINO, as well as HAT, constitute competing predictions of what is to come.22

The items in competition can of course be affected by context. For instance, if /kæp/ in the input is preceded by /wɛər/ (“wear a”), cap will be semantically primed, at the expense of other competitors and their predictions. And, alternatively, if /kæp/ turns out to be followed by /tən/, captain will be further activated and the other candidates and their predictions will be suppressed. In short, within-item pre-activation begins the chain of events that produce cohort effects in word recognition.

A second case is closer to the sort of example often employed in studies of prediction (e.g. Kuperberg & Jaeger, 2016; Morgan & Levy, 2016). Suppose one hears the beginning of a sentence:

(9) She put salt and …

This fragment invites the prediction that the continuation will be something like (10a), and not something like (10b,c,d) or many other possibilities.

(10) a. She put salt and pepper on the eggs.
    b. She put salt and mustard on the shopping list.
    c. She put salt and a great deal of pepper in the soup.
    d. She put salt and, come to think of it, way too much pepper in the soup.

This prediction is driven by the prominence of the stored collocation salt and pepper, which has linked semantic, syntactic, and phonological structures.24 The collocation has relational links to the entries of the individual words salt, pepper, and and, as shown by the subscripting in Fig. 6. (And is treated as a schema that joins variable conjuncts.)

Fig. 6 shows the flow of activation beginning at the moment when (9) is heard. The input activates the individual phonological words salt and and, which pass activation “bottom-up” to their respective semantics. But in addition, the whole collocation salt and pepper is activated through its phonology, even though pepper has not been heard. In other words, the string salt and … serves as an incipit and pepper is its remainder, hence a prediction in our sense.

However, there is a complicating factor: at the same time, the semantics of salt activates that of pepper through ordinary spreading activation (the dashed arrow in Fig. 6). If that is the case, PEPPER should send activation down through the interface link to the phonology /prəp/. But then, if indeed pepper is pre-activated, might this be accounted for simply by ordinary semantic priming? Is there any need to appeal to within-item pre-activation?

There is. It seems fair to assume that the semantic relational link between salt and pepper is equipotential, that is, all else being equal, pepper can prime salt to the same degree as salt primes pepper. If semantic priming is all that is going on in Fig. 6, then pepper and salt ought to be about as frequent as salt and pepper – speakers should not be biased to say one rather than the other, given their semantic equivalence. But this is not the case. A COCA search yields 5890 hits for salt and pepper vs. only 86 for pepper and salt. The disparity suggests that the semantic priming of pepper by salt is not doing much work, compared to the within-item pre-activation by the collocation. Moreover, there are 10,269 hits for salt and, suggesting that this incipit correctly predicts pepper more than half the time, again attesting to the relative activation strength of the collocation.

To dwell on this point a little more, consider pairs of color names. Black and white and black and blue are fixed expressions; they have 8450 and 425 hits in COCA respectively; while their reversals, white and black and blue and black, have 974 and 128 respectively, paralleling the disparities with salt and pepper. In contrast, where there is no prominent fixed expression, there is far less disparity. For instance, black and green has 92 hits, while green and black has 136; blue and green has 407 hits, while green and blue has 338. The collocation Black Friday, neither word of which primes the other, has 1986 hits; the non-collocation Black Thursday has 23. For a different sort of example, dogs and cats, with 719

---

22 We do not address what has been called prediction in language production, which we would prefer to call planning. Nor do we say anything about prediction in the motor control of speech (e.g. Hickok, Houde, & Rong, 2011). We also do not discuss the many processes that invoke post- or retrodiction, a.k.a. right-context effects. These are situations in which part of an input is indistinct or ambiguous, and subsequent input clarifies the intended message. Such phenomena include for example syntactic and semantic disambiguation, the phoneme restoration effect (Leonard, Baud, Sjerp, & Chang, 2016; Samuel, 1981; Warren, 1970), and noisy channel effects (Gibson et al., 2013; Levy, 2008). We believe our approach can be extended to these phenomena, but we will not attempt it here (but see Jackendoff, 2007 on the late disambiguation of It’s not apparent vs. It’s not a parent).

24 Note that the structure of salt and pepper is completely redundant. Nevertheless, it is stored in the extended lexicon, with all its structure, because this is a fixed phrase that is known as such by speakers of English, in contrast with a novel expression such as chairs and rugs, which is a product of free generation.
Cognition 224 (2022) 105050

10

5.3. What is the scope (the possible extent) of predictions?

Much of the discussion of prediction in the literature is occupied with predicting the next word in the sentence, for instance pepper in (9) (Huettig, 2015; Pickering & Gambi, 2018, for reviews). However, this is only the tip of the iceberg.

Consider what is happening earlier in the processing of example (9), at the moment when only She put ... is heard. Here, the very next word cannot realistically be predicted. However, the verb put does predict that it will be followed by an NP and a locative PP, whatever words these happen to consist of. This syntactic prediction follows from the lexical entry of put, which looks roughly like (11). Its syntax stipulates a variable direct object, linked to a variable entity in motion in the semantics, followed by a variable locative PP, linked to an endpoint of motion. (These variables are the Parallel Architecture’s notation for subcategorization features in syntax and selectional restrictions in semantics. The details of coinexi”n are not too critical for the present context.)

In other words, this is a case of within-item prediction that is not about predicting the next word. When the phonological word /put/ is encountered in the input, it activates not just the verb pronounced /pʊt/, which functions as an incipit, but also the syntactic variables NP and PP, which together function as the remainder. But there is no phonological prediction; the particular words that make up the NP and PP remain unspecified.

To see the effect of this syntactic prediction, suppose the input turns out to be She put salt yesterday. Then the predicted locative PP is absent, so the predicted remainder conflicts with the actual input yesterday. Furthermore, the verb’s need for a locative PP is non-negotiable, so the sentence ends up being judged ungrammatical.

A slightly different input, She put salt on ... raises semantic expectations that go beyond the choice of the next word. First, the tense of put sets up a semantic expectation that any time expressions to come will refer to the past, so that She put salt on her eggs tomorrow will occasion problems on the final word. Second, the object of the PP is expected to denote some sort of food, in particular something that one would put salt on. Hence eggs is less of a surprise than, say, éclair, which in turn is less of a surprise than socks (as in the famous example of Kutas & Hillyard, 1980a, 1980b). Unlike the salt and pepper case, this case does not invoke particular collocations, so they are (presumably) cases of pre-activation by semantic association, again with no accompanying phonological prediction.

This same input, She put salt on ... can create yet another kind of semantic violation, seen in (10b). The canonical use of put denotes caused motion of a physical object to a physical location. Thus it strongly predicts that the object of on will be the surface of a physical object – if not food, then something like a table or shelf. But suppose the input continues ... on the shopping list. Then the agent is not causing salt to move – in fact she is not acting on salt at all. Rather, she is creating an orthographic representation of the word salt on a list, a wholesale reconstruction of the meaning of the sentence, thwarting the canonical prediction. (There is of course also a physical reading as well: sprinkling salt on the piece of paper on which the list is written. But this is
Thus like the case of put, this case is an instance of within-item prediction, except that it is at the level of segmental phonology. It should be noted, though, that such within-item phonological prediction is typically much weaker than semantic prediction (Nieuwland et al., 2018). This is presumably because semantic pre-activation can add up as the gist of a sentence develops. For instance, the meaning of the second word can bias the choice of the seventh word, as in She drank some unusual sort of..., whose probable continuation is a word denoting a drinkable liquid. In contrast, phonology does not have similar cumulative effects: what the second word sounds like has little effect on what the seventh word sounds like.

Turning back to prediction of syntactic structure (many cases of which are discussed by Ferreira & Qiu, 2021), The determiner the activates the NP schema (13) (= (6)), of which it is the leftmost part.

Thus Det serves as an incipit and predicts a noun to come, which will be the determiner’s head. But it does not predict which noun,28 nor what adjectives and other modifiers might intervene, especially in extended examples such as a highly unusual and insufficiently tested technique. Similarly, hearing if predicts first a clause of undetermined length, optionally followed by then, plus another clause. This is in line with experimental evidence that readers who have previously read an either-clause will predict an or-clause to follow (Staub & Clifton, 2006).

For another case, a wh-word at the beginning of a question or relative clause predicts a syntactic gap to follow, but it does not determine where that gap will occur. Rather, the wh-word serves as an incipit to the extended construction that licenses long-distance dependencies (Chaves & Putnam, 2020; Culicover & Jackendoff, 2005; Sag, 2010). In quite a different realm, poetic conventions predict metrical patterns and the placement of rhymes – while making no commitment whatsoever to the choice of words.

Pushing the pre-activation of syntactic structure a bit further, we arrive at a natural account of so-called left-corner parsing (Resnick, 1992). Suppose an utterance begins with the input /ðə/. As suggested above, this pre-activates the NP schema (13). NP in turn pre-activates (14a), the treelet for S, of which it is the left-hand part. The result is (14b). The VP in (14b) pre-activates its parts, in particular an initial V, as in (14c). The result is (14d), all of which except for the determiner is the result of within-item pre-activation. Hence from the minimal input the...

...the processor can anticipate (or predict) a significant chunk of structure to come.29

(14) a. Syntax: [s NP <Aux> VP]
   Phono[y: /ðə/1]
   b. Syntax: [s [VP Det1 <A> N ...] <Aux> VP]
   Phono[y: /ðə/1]
   c. Syntax: [vp V ...]
   d. Syntax: [s [VP Det1 <A> N ...] <Aux> [vp V ...]]
   Phono[y: /ðə/1]

Of course the anticipated structure in (14d) is not always correct. Suppose the second word is more. This brings into play the lower-frequency schema for the comparative correlative construction (e.g. the more I read, the less I understand: Culicover & Jackendoff, 2005), which then competes with the canonical NP structure. The competition can go on for several words, as in (15a,b), without any noticeable sense of garden-pathing.

(15) a. The more appealing and insanely popular theory isn’t always correct. (comparative NP]
   b. The more appealing and insanely popular a theory is, the more likely that it’s incorrect.
   [top comparative correlative construction]

It might be predicted that the processing load in these examples increases as the ambiguity continues. We are not aware of any experimental investigations of this particular situation, but it falls out as a natural consequence of the PA’s account of syntactic parsing. We welcome an experiment.

The upshot is that prediction in the sense proposed here appears at all scales, from the next word to both smaller and larger structures, and to both relatively concrete and more abstract structures. This conclusion is supported by numerous experimental studies that provide evidence for pre-activation of various parts of phonological structure (e.g. Nieuwland, 2018; Pickering & Gambi, 2018). There is also strong evidence consistent with pre-activation of semantic representations (e.g. Altmann & Kamide, 1999, Federmeier & Kutas, 1999; Federmeier, McLennan, De Ochoa, & Kutas, 2002; Mani & Huettig, 2012; and many others) as well as orthographic representations (Laszlo & Federmeier, 2009).

5.4. Further considerations

To summarize: the generation of predictions via pre-activation is a natural byproduct of accessing stored linguistic representations. Consequently, prediction occurs constantly in language use, as claimed, for instance, in Federmeier, 2007; Ferreira & Chantavarin, 2018; and Huettig, 2015. Predictions are an inevitable consequence of a richly connected lexicon and a language processor with a robust and efficient system of lexical access.

To push the point further, there are many situations where there is nothing much to base prediction on. Consider a sentence-initial incipit Today the... In the absence of contextual pre-activation, there is no highly pre-activated candidate for the next word; the only prediction is that the is the beginning of an NP of completely indeterminate content. Hence, nothing either reinforces or interferes with processing of the next word(s). Similarly for the case mentioned earlier, She put... The pre-activation theory therefore takes it that people do indeed often predict words to come, but only when there are predictions to be made, i.e. when there are one or more reasonably highly pre-activated candidates.

The bulk of published experimental evidence on prediction in
language may well give the impression that language processing is always predictive, in conflict with our conclusion here. However, it is not obvious that examples such as *Today the... or She put... are any less common than the strongly predictive contexts that often form the experimental paradigms in the literature (Huettig & Mani, 2016). Many experimental studies have used experimental sentences with extremely high cloze probabilities, and then have generalized their conclusions to all of language processing. This bias in the empirical literature is further exacerbated by the well-known prejudice against publishing null effects (Van Assen, van Aert, Nuijten, & Wicherts, 2014) and replication failures (Nieuwland, 2018; Nieuwland et al., 2018). Recent studies that have directly investigated prediction with more diverse stimuli, i.e. varying their predictability (Frisson, Harvey, & Staub, 2017; Huettig & Guerra, 2019; Luke & Christianson, 2016), appear very much consistent with the present account. More generally, we believe that studies using experimental stimuli across the predictability range (e.g. Brothers & Kuperberg, 2021; Ratas & Hillyard, 1984; Luke & Christianson, 2016; Smith & Levy, 2013) will prove most informative.

6. How does pre-activation theory relate to probabilistic accounts of prediction in language processing?

As promised in section 1, our PA-based theory of prediction via pre-activation has much in common with other approaches to prediction. In particular, the literature offers a variety of influential theories based on Bayesian and related probabilistic and information-theoretic frameworks (Fale, 2001; Levy, 2008, Kuperberg & Jaeger, 2016; Norris et al., 2016; cf. also Attneave, 1959; Shannon, 1948). These accounts are closely related to earlier probabilistic (Jurafsky, 1996) and constraint-satisfaction theories of sentence parsing (MacDonald et al., 1994; Trueswell et al., 1993). Common to these approaches is that a prediction is taken to come with a probability — the processor is hedging its bets. Bayesian statistical decision theory proposes that language processing is “optimal” or “rational” and that people are “ideal observers”. In a similar vein, information-theoretic approaches assess the informativeness (or surprisal) of a linguistic phenomenon, given a hearer who is ideally equipped to receive the information. According to both these approaches, language users make use of fine-grained probabilistic knowledge from their past language experience to compute fine-grained probabilistic predictions about current and upcoming linguistic structures.

The PA-based theory is on the whole intertranslatable with the general tenets of Bayesian and related probabilistic and information-theoretic accounts. Many probabilistic predictive effects in language comprehension, we conjecture, stem from features of linguistic structure, in particular, PA’s notion of an “extended lexicon”. Prediction in our theory is conceptualized in terms of multiple possible continuations of the incipient, each competing to be “what is heard.” Within this framing, the counterpart of the probability of a candidate continuation is its relative activation strength compared to that of competing candidate remainders. This can be reinterpreted as a probability by normalizing over the total degree of activation among current candidates.

The counterpart of surprisal in the present approach is in the interaction between pre-activation and subsequent input. A continuation of the input that is typically described as low-surprisal is one that matches a highly pre-activated remainder (i.e. it is strongly predicted), and whose processing is therefore enhanced. A continuation that is typically described as high-surprisal matches only a weakly pre-activated remainder or none at all (i.e. it is predicted weakly at best); hence a strong competing prediction will interfere with processing. If there is no strong alternative prediction, as in *Today the..., processing will be neither enhanced nor inhibited.

Given this degree of equivalence between the pre-activation account and a more standard probabilistic approach, one might wonder what is to be gained by adopting an account couched in terms of pre-activation. A first strength of our account is the Parallel Architecture’s explicit description of linguistic structure and language prediction at all levels of representation (semantics, syntax, phonology, and the relationships between them). Our account extends to all scales, from within words to long-distance dependencies and other abstract syntactic structures (comparable to Levy, 2008; MacDonald et al., 1994; Trueswell et al., 1993; which assume that multiple individual levels of representation can converge as a lexical “bottleneck”; and to Kuperberg & Jaeger, 2016, who attempt to sketch out a hierarchical generative architecture in which information is treated “probabilistically at all levels of representation at once.”).

Second, the pre-activation approach proposes that prediction is modulated not only by within-item pre-activation, but also by the resting activation of the items in question, by identity and associative priming, and by the activity of other candidates in competition for — or in support of — “what is heard”. The explicit description of linguistic structure permits the theory to recognize this plurality of sources. At the same time, they can all be regarded as affecting the moment-to-moment degree of activation. (Kuperberg & Jaeger, 2016, and Smith & Levy, 2013, offer similar accounts in probabilistic terms.)

Third, the Parallel Architecture can be considered a realization of Marr’s (1982) computational level of analysis, in that it describes the mental representations that the language faculty computes and stores. However, the pre-activation theory of prediction arguably goes beyond the computational level by focusing on how linguistic representations are used to generate predictions online. It therefore lays the groundwork for integrating the computational level with Marr’s algorithmic level, describing in detail all the steps in how predictive language processing “is done.”

It should be acknowledged that any theory of predictions is subject to a crucial methodological constraint: for the most part, a prediction can only be detected by virtue of its effects on subsequent input, such as cloze probabilities, lexical decision, reading times, ERP components, and eye movements. In order to conclusively demonstrate the presence of pre-activation/prediction, it is necessary to measure its effects at the moment when an incipient and its remainder(s) have been activated, but the input has not yet presented a continuation that can interact with the predicted remainder(s). Using EEG and MEG techniques, Wang et al. (2020) succeed in detecting predictions of animate vs. inanimate – though not particular words – prior to the onset of the continuation of the input. (For further discussion see Baggio & Hagoort, 2011; DeLong et al., 2005; Huettig, 2015; Mantegna, Hintz, Ostarek, Alday, & Huettig, 2019; Nieuwland et al., 2018, 2020; Pickering & Gambi, 2018.) It is here that computational modeling might well prove useful, allowing one, as it were, to step inside the simulation of the machine, as activation strength and timing are manipulated. This is a major task for future research.

7. Conclusion and further directions

We have argued that the Parallel Architecture’s *pre-activation theory* is a linguistically and psychologically plausible theory of prediction in language processing. It takes seriously the question of the form of predictions and how they are generated, and it shows how the generation of predictions can be considered a natural byproduct of the use of these representations in processing. The nature of within-item prediction follows from the fact that language processing is constantly accessing words, multi-word units, and abstract schemas. In response to a piece of input, it generates multiple structures on multiple levels and at

---

30 This problematic situation is likely to improve, given recent developments and suggestions made by the open science movement (Zwaan, Ets, Lucas, & Donnellan, 2018).

31 We note that Marr’s algorithmic level does not necessarily involve literal algorithms. We are interpreting it as a theory of processing, and we believe this is the sense that Marr intended.
different scales. In our terms, an incipit activated by input activates one or more corresponding remainders. These constitute predictions of what is to come. In other words, a prediction amounts to the as yet unheard part of a lexical item whose beginning has been activated by the input. Remainders that are simultaneously in play vary (a) in the degree of reactivation of the items of which they are a part, (b) in the strength of pre-activation, including identity, semantic, and contextual priming, and (c) in the number and strength of competing remainders.

The main innovations in the pre-activation theory come from innovations in the linguistic theory. Most important have been (a) that the Parallel Architecture’s extended lexicon contains not just words but collocations, idioms, meaningful constructions, and schemas, the latter of which take the place of traditional rules, and (b) that activation spreads in structured fashion, following the pathways of PA’s interface links and relational links. These innovations lead to a sharpened theory of processing, in which the necessary steps in processing can be identified, and from which the pre-activation theory falls out as a natural consequence. An important feature of the theory is that it unifies cohort effects (e.g. captain), word and collocation prediction (salt and ...), and structural prediction (if S, then S).

The pre-activation theory offers a natural but at the same time representationally explicit account of processing, including such effects as pre-activation through spreading activation, the lexical boost effect, the inverse priming effect, and priming of all sorts, including structural priming and semantic priming. The theory in turn is based on the Parallel Architecture theory of linguistic representations, which gives equal weight to phonology, (morpho)syntax, and semantics, and which encodes linguistic units of all sizes and all degrees of abstractness in a common format in the extended lexicon. Thus it fits comfortably with linguistic and psychological theorizing as well as with empirical evidence.

CRediT authorship contribution statement

Falk Huettig: Conceptualization, Writing – review & editing. Jenny Audring: Conceptualization, Writing – review & editing. Ray Jackendoff: Conceptualization, Writing – review & editing.

Acknowledgments

We are first of all grateful to the 2019 Norwegian National Graduate School in Linguistics, which took place in exotic Svalbard. FH and RJ, who had not previously met, were on the faculty; and one bright sunny night, drinking below decks during a chilly and largely fruitless whalewatch, they found common ground that inspired the present paper.

We wish to thank Dagmar Divjak, Fernanda Ferreira, Hartmut Fitz, David Gow, Rob Hartsuiker, Greg Hickok, Sahil Luthra, Jim Magnuson, Luca Onnis, and Eva Wittenberg for forbiddingly detailed comments on our first draft. We thank Gary Dell and two anonymous reviewers for their comments on a previous version of this paper. Thanks also to Neil Cohn for the illustration in section 4.

References


F. Huettig et al.

Cognition 224 (2022) 105050


