Published in final edited form as:

Trends Cogn Sci. 2023 November ; 27(11): 1032–1052. doi:10.1016/j.tics.2023.08.003.

Prediction during language comprehension: what is next?

Rachel Ryskin[1,*], **Mante S. Nieuwland**[2,3]

[1]Department of Cognitive and Information Sciences, University of California Merced, 5200 Lake Road, Merced, CA 95343, USA

[2]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

[3]Donders Institute for Brain, Cognition, and Behaviour, Nijmegen, The Netherlands

Abstract

Prediction is often regarded as an integral aspect of incremental language comprehension, but little is known about the cognitive architectures and mechanisms that support it. We review studies showing that listeners and readers use all manner of contextual information to generate multifaceted predictions about upcoming input. The nature of these predictions may vary between individuals owing to differences in language experience, among other factors. We then turn to unresolved questions which may guide the search for the underlying mechanisms. (i) Is prediction essential to language processing or an optional strategy? (ii) Are predictions generated from within the language system or by domain-general processes? (iii) What is the relationship between prediction and memory? (iv) Does prediction in comprehension require simulation via the production system? We discuss promising directions for making progress in answering these questions and for developing a mechanistic understanding of prediction in language.

Why predict?

As you read these words, you are probably guessing which words might come subsequently (were you surprised that the last sentence did not end with ‘next’?). Similarly, in most everyday conversations, as you listen to another person, you are processing what they are saying and planning your own response. Despite the typically rapid pace of speech – around 2–3 words per second – and the many things you may want to consider before responding, you will start speaking within about a quarter of a second after they finish, if not sooner [1,2]. How do you achieve this impressive feat? This simple question belies the complexity of the cognitive processes involved and the fact that, at its core, it is a question about the fundamental workings of the language system in the human mind and brain. An increasingly popular hypothesis is that people are generally able to keep up with language input by predicting what comes next – by activating the meaning and potentially other aspects of words ahead of time. This hypothesis was somewhat controversial at the turn of the past century, but research over the past two decades has demonstrated the psychological reality of

*Correspondence: rryskin@ucmerced.edu (R. Ryskin).

Declaration of interests

The authors declare no conflicts of interest.

'linguistic prediction' beyond any reasonable doubt. The limelight has therefore shifted from initial existence proofs to the underlying cognitive architecture and mechanism.

Researchers attempting to simulate language learning and use have often found that the goal of predicting the next word (or other linguistic unit) appears to endow models with many desirable properties, including the ability to learn to generate human-like linguistic sequences implicitly from language input and the ability to replicate neural and behavioral patterns of humans engaged in language comprehension [3–6]. In addition, predictive processing has been a highly productive explanatory framework across domains of cognitive science [7,8]. In the domain of vision, in particular, the computational-level framework has been linked to evidence for a particular instantiation of probabilistic prediction at the mechanistic level (used here to refer to algorithmic and/or implementational levels in Marr's terms [9]) – predictive coding [10] (Box 1 and Figure 1). According to the most typical version of predictive coding, prediction signals travel down to the lowest levels of representation, whereas signals carrying prediction error resulting from the comparison of sensory input and predictions travel upward along parallel pathways. Application of this framework to the domain of language holds promise. However, prediction at the computational level need not be implemented by a mechanism with an explicitly predictive objective. Models with different objective functions (e.g., incremental word recognition, homeostasis) can also account for some aspects of predictive behavior in humans [11,12].

At the computational level of analysis, many proposals contend that humans predict upcoming linguistic input based on internal models of the environment, and update those models based on some comparison between the prediction and the received input ([13,14] *inter alia*). However, accounts differ in terms of the specifics of these computations, and the landscape of mechanistic proposals is muddled. In the current review we first synthesize what is known about the predictive computations of the human language comprehension system – what information is used as input to the computational system and what makes up the output, as well as how variability can be understood in terms of additional constraints impinging upon the system. We then turn to questions about how these computations may be implemented at the process level, emphasize recent developments and unresolved debates, and look ahead to how the field might move toward a mechanistic understanding of linguistic prediction. Throughout, we focus on high-level language comprehension, typically (but not exclusively) at the sentence level, and set aside detailed discussion of lower-level auditory or visual feature prediction.

Prediction at the computational level

Decades of psycholinguistic research have shown that the processing of a word depends in part on the words that preceded it – the sentential context. When a word (e.g., 'next') is highly expected given the preceding context ('Which word will come … ?'), it is identified more easily [15], it is read more quickly [16], and it elicits a smaller (negative) electrophysiological response in the 'N400' time-window [17] than when it is not expected given the context (e.g., 'The following is an example of a word, … '). Moreover, listeners direct their attention toward images that they expect to be referred to next as they hear a sentence unfold [18,19].

Attempts to formalize the notion of 'predictability' or 'expectation' provide converging evidence regarding the core role of prediction in language processing. The probability of a word in context (i.e., given the preceding words, syntactic structures, etc.) can be estimated using corpus counts, language models, or cloze tasks; words that are lower in contextual probability are read more slowly [20] and elicit more negative N400 responses [21,22]. However, whether processing difficulty is a log-linear function of probability in context [23], as stipulated in surprisal theory [24], is still debated [25–28]. Similarly, encoding models predicting neural activity (fMRI recordings) perform better when using representations that incorporate context (via a neural network) relative to representations of linguistic input that do not (word embeddings from a distributional semantic model) [29]. (Box 2 and Box 3 provide an overview of the methods used to investigate linguistic prediction and how they reflect different aspects of language processing).

The interpretation of reading time and event-related potential (ERP) findings as evidence for prediction was originally the topic of some debate: most behavioral and neural measures do not provide a read-out of the prediction itself but instead provide a measure of how new input is processed in light of what may have been predicted. As a result, it is difficult to distinguish between a case where a listener's processing of a word is facilitated because it had been predicted before the bottom-up input was received and a case where the received input is integrated with the preceding context and this integration is facilitated when the input matches the context. However, other findings provide compelling evidence that integration cannot explain all apparent prediction effects [30–33], although it may also play a role [34].

Moreover, multivariate approaches to analyzing neural data have begun to reveal the predictive processes occurring in the moments before the crucial input is received. For instance, representational similarity analysis (RSA [35]) indicates that, before the crucial input, neural patterns corresponding to predictions of the same words (in different contexts) are more similar than the neural patterns corresponding to predictions of different words [36], and these neural patterns could contain coarse-grained semantic information such as the animacy of the upcoming word [37]. Similarly, neural patterns preceding a target word may be more similar to those evoked by the target word itself when that word has high contextual predictability [38]. Furthermore, vector-space representations of words (GloVe embeddings) show that information about a predicted word is encoded in the neural activity which closely precedes its onset [5]. In mechanism-agnostic terms, we can conceive of prediction as the mental state of the comprehender fashioning itself to resemble what is likely to come next, such that the more predictable the bottom-up input from the context, the less change of state takes place when the input is received and processed.

The input: the context for prediction

The human language system appears to predict on the basis of a constellation of input properties (reviewed in [14,39]). To make estimating contextual probabilities tractable, the context is often operationalized as the preceding words in a sentence and their semantic and syntactic properties. Phonological, lexical, and syntactic sequence-based contextual probabilities appear to modulate neural responses during naturalistic listening [40], as

do those derived from models with hierarchical structure [41]. However, these sequence-based estimates may not fully capture human predictability judgments [42]. For instance, the semantic similarity (e.g., derived from distributional semantics models) of a word and its context seems to explain additional variance in neural signals above and beyond contextual probabilities [43,44], although recent work shows that this effect is minimized when predictability is estimated using large language models with a more sophisticated representation of the preceding context [45].

In fact, comprehenders consider a broad scope of contextual information. For instance, the larger narrative within which a sentence is embedded and world knowledge both affect what is predicted [46,47]. Pragmatic cues and inferences about the intentions of the speaker also contribute to predictions. Listeners predict how speakers will correct themselves if they misspeak (e.g., listeners look to a dog if they hear the speaker say '… his cat, uh I mean his …' [48]). They also consider the idiosyncrasies of the speaker and adapt their predictions accordingly. For instance, listeners take into account what the speaker is likely to talk about given their preferences and interests [49], how fluent the speaker is [50], whether they are a 'native' speaker [51], and whether the speaker has a proclivity for surprising sentence completions [52]. In this regard, people sometimes take into account statistical regularities in the context (e.g., if predictable endings are less likely than unpredictable endings across trials in an experiment) when predicting upcoming words, although they may not always do so [53–57].

The output: the content of prediction

What information is carried by predictions, and at what grain size? The evidence for semantic or conceptual prediction is robust (reviewed in [58]). Instead of consisting of a single maximally probable word, predicted information appears to be graded and, consequently, encompasses many aspects of form and meaning [30,59]. Depending on how the neural responses are carved up, the predictions of comprehenders can be shown to capture coarse-grain features such as animacy [37], as well as finer-grained features that uniquely identify a lexical item, when the context is constraining [36].

Recent work using naturalistic paradigms – where participants listen to audio recordings (e.g., of stories, talks, etc.) without any explicit task or experimental manipulation – supports a role for hierarchical prediction across multiple representational domains (semantics, syntax, phonology) and timescales, where higher-level information constrains predictions at lower levels [60–63]. For instance, acoustic features of words appear to be more sharply encoded by the brain when those words are semantically related to their context [64]. Moreover, when uncertainty about an upcoming word is high, fast-timescale (4–10 Hz) responses in primary auditory cortex are stronger than when the uncertainty before the word is low, suggesting that sensory sampling is increased when uncertainty is higher [65]. Slower-timescale (0.5–4 Hz) responses, outside primary auditory cortex, are increased when a word is more surprising in context, suggesting that these responses reflect updating of the internal model at higher levels of representation (*cf* [66]).

Although this work suggests that even low-level speech perception is facilitated by prediction, these naturalistic studies often cannot temporally disentangle the consequences

of prediction from prediction itself. Whether top-down predictive signals reach sensory-perceptual (visual or auditory) representations has been the focus of a long-standing debate in this literature. In one well-studied paradigm, ERP evidence was used to argue that listeners not only predict the upcoming content word/meaning in a sentence such as 'The boy went outside to fly … ' but they also predict the form of the article that would need to precede it (i.e., they expect 'a' before 'kite', but 'an' before 'airplane' [67]). Crucially, recent replications and reanalyses suggest that the effects of specific form predictions may be much subtler than was initially assumed [68–71] and that careful consideration of the real-world probabilistic relationships between articles and nouns may reveal a more nuanced picture of these prediction effects [72–74]. Similarly, listeners can predict phonological features of an upcoming word when reading rhyming text [31]. However, these experiments do not address to what extent predictions reach lower levels of perceptual representations in typical comprehension [75].

Constraints and variability in prediction

What are the constraints on linguistic prediction? According to predictive processing/Bayesian brain accounts, language users continuously consider/update the full probability distribution over all possible linguistic inputs. However, in its unconstrained form, this proposal is computationally intractable [76,77]. Resource-rational accounts propose that the mind and brain perform rational inferences within the constraints imposed by the biological and informational limits of human brains [78], but what these are, in particular with respect to language, remains to be determined. Investigating variability between people in terms of their predictive processing abilities/tendencies provides some initial clues about these limits [79]. In particular, three populations have often been investigated in this context because prediction appears to be systematically different in these individuals relative to the default comparison group of young adults: children, second-language learners, and older adults.

Implicit in the predictive processing framework is the idea that humans have an internal model of the world from which their predictions are derived. This model cannot be innate and must therefore be the result of learning. One common view, inspired by connectionist models, is that prediction is not an end in itself, but is instead a means by which the internal model can continuously adapt to more accurately reflect the world [3,80]. This view is supported by evidence that prediction updates in young children (as measured by eye movements or ERPs in response to an unexpected word) are larger when children have more vocabulary knowledge [81–83]. By the same token, predictions become more accurate as children refine their model of the language [84–86]. An alternative view suggests that the flow of causality is reversed: once children have sufficient language experience, they begin to predict [87]. A closely related view proposes that general cognitive maturation explains the delay in adult-like language prediction. The executive resources and working-memory skills of children increase until their mid-20s [88,89]. On the assumption that prediction involves such executive resources (discussed in further detail below), the ability of children to predict upcoming linguistic material develops in tandem.

Similarly, adult language learners who have less (or a qualitatively different) experience with the language in which they are tested (e.g., second-language learners, individuals

with low literacy skills) also appear to predict less than their same-age counterparts who have more experience with the language [90–93]. As with the development of prediction in children, these differences may be explained by the lower accuracy of the prediction-generating internal model of these individuals [94]. Indeed, these differences go away when the second-language learners are highly proficient in the language in which they are being tested [95,96]. Alternatively, the population difference may be mediated by executive resources. In particular, comprehenders who are less fluent in the language may use up their executive resources on basic aspects of incremental comprehension (e.g., word recognition), leaving few resources available to generate predictions [96,97]. In other words, differences in predictive processing may result from differences in other aspects of comprehension.

Finally, the difference in neural or behavioral responses between predictable and unpredictable stimuli is typically reduced for older adults relative to their younger counterparts, suggesting that the former are less successful in using context to predict ([98–101], *cf* [102,103]). As with other populations, one proposed explanation of this apparent decline in prediction is that it is mediated by executive resources [104], which are known to decline from the 20s into older adulthood [89]. By contrast, age-related declines in prediction effects may be driven in large part by experience [105,106]. As we age, the brain refines and optimizes its internal model of the world, and this may lead to more efficient prediction and attenuated updating in the face of novel sensory input [107]. Older adults have more experience with the language overall, as evidenced by larger vocabularies [108], and the nature of their linguistic exposure may differ substantially from that of younger adults (e.g., different sources, more familiarity with outdated terms or constructions). In sum, the content of the predictions of older adults may differ from the predictions generated by younger adults owing to differences in their internal models of the language. More broadly, uncovering how the computation of prediction is constrained by human resource limits will require knowing what the relevant resources are, and to achieve that we will need to understand the mechanisms which implement prediction.

Toward a mechanism for prediction in language

Although some aspects of the computations continue to be debated and refined, the evidence is compelling that humans learn about the properties of the world and the language, and can use this internal model to generate probabilistic predictions about upcoming linguistic inputs (Figure 2). However, the landscape of possible mechanisms underlying these computations is vast, and, in our opinion, exploring it is the next frontier for this field (Box 4 for a discussion of one approach). In what follows we first briefly summarize a few mechanistic proposals which have started to emerge, and we then discuss key unresolved questions which may help to constrain the space of possible mechanisms and guide future research efforts (summarized in Figure 3).

Some current proposals

Predictive coding—Predictive coding has become a dominant proposal for how prediction in various domains of perception and cognition may occur at the level of algorithm and/or implementation (Box 1 for a brief overview of predictive coding outside

the domain of language). A predictive coding account of language posits that linguistic predictions from higher levels (e.g., meaning/syntax) are sent down to lower levels (e.g., word form/speech perception), and a measure of mismatch between the input and prediction (i.e., errors) is sent back up the hierarchy. Predictive coding may be a fruitful way to instantiate the computational-level proposal of a hierarchical generative framework for language comprehension [14,109]. Another proposal [110] explicitly connects the amplitudes of N400 ERP components to precision-weighted bottom-up predictive coding error signals. Indeed, a recent attempt to implement predictive coding as a mechanistic account of high-level language comprehension shows promising correspondences between the timecourse of prediction errors and the timecourse of the N400 across multiple contexts [111].

Error-based learning—Although error signals in predictive coding are part of the information flow between levels of the hierarchy that constitutes inference or comprehension, traditional connectionist/neural network approaches to language have often incorporated a prediction error primarily as the driver of learning [3,13,112]. In these architectures, the model attempts to predict the next element in a sequence and, upon receiving the next input, computes the error between its prediction and the input. This error is then propagated back to update the model weights, with the goal of minimizing future (average) prediction error. The error signals themselves, or other values derived from the networks, can be tied to behavioral or neural indices of prediction. For instance, one account [6] relates the magnitudes of prediction errors from semantic and sequencing layers of a neural network to the amplitudes of the N400 and P600 ERP components, respectively. However, these models have primarily been tested on small toy datasets, and the relative merits of different variants (e.g., in terms of fit to data as well as neurobiological plausibility) have not yet been systematically evaluated.

Alternatives—There are numerous other possible model architectures. Many of these may not incorporate explicit prediction (or prediction error) but may still be able to account for the empirical data patterns reviewed in the previous section. For instance, neural networks trained to simulate incremental word recognition can display behaviors consistent with predictive processing, notably reductions in signal strength when an input is predictable [12,113]. Further, networks of connected reservoirs which receive input from the environment and simply strive to maintain homeostasis display predictive processing-like behaviors: their spiking activity patterns in response to partial inputs are most similar to those of the most likely continuation given the training data [11]. More comprehensive testing will be necessary to determine whether these mechanisms succeed in accounting for the full scope of predictive processing effects that have been documented and described at the computational level.

Key unresolved questions

Is prediction essential or optional?—Many current frameworks in the cognitive science of language view the prediction of upcoming linguistic content as part-and-parcel of the functioning of the human language system (e.g., [13,14,23,39,114]). However, on other

accounts, prediction is a beneficial but optional skill [115], one that may emerge as the result of literacy [116].

According to accounts that view prediction as an essential computation, an entirely passive comprehension mode is unlikely for adults reading or listening to typical sentences. This view draws on support from the myriad findings (in the section on Prediction at the computational level) of spontaneous predictive behavior during language comprehension in naturalistic settings, as well as from evidence of neural signals of prediction generation before the bottom-up input is received (e.g., [5,36,38]). In addition, a constraining sentence context elicits a more negative slow-wave response ('semantic readiness potential') preceding the crucial word, and this potential appears to have different spatial topographies for sentences describing hand-related versus face-related actions [117], suggesting that it encodes some aspect of the semantic content of the context and/or prediction. Similarly, the N400 amplitude in response to adverbs (e.g., 'often') is greater when the adverb increases the predictability of the target word that follows it [118], consistent with the view that the N400 indexes the updating of a probabilistic meaning representation [105].

By contrast, proponents of the view of prediction as an 'optional strategy' point to studies showing variability across populations (in the section on Constraints and variability in prediction) and the apparent absence of prediction effects under 'adverse' conditions. For instance, in a visual-world paradigm (VWP), when the time-window for prediction is short and minimal preview time is provided, predictive fixations to a target noun that is gender-congruent with the article are reduced relative to slower listening conditions [119] (note that there is an important distinction to be made between scenarios where humans refrain from predicting when input is rapid and scenarios where researchers fail to detect predictions on faster timescales: current measurement tools, even those with fast temporal resolution, often cannot distinguish between these scenarios). Similarly, reading-time differences between predictable and unpredictable sentence completions approach zero when the proportion of sentences with unpredictable endings in the experiment is high [120].

One challenge in disentangling these two perspectives is that an apparent lack of difference (in eye movements, ERPs, etc.) between a predictable and an unpredictable sentence continuation is difficult to interpret. In particular, experimental manipulations of predictability vary in terms of their effect sizes. In some cases a real effect of predictability may be so small as to require very large amounts of data to reliably detect it (discussed in [69,70]). Setting aside methodological concerns, this could mean that (i) the listener predicted nothing and simply passively waited for the next input for both types of sentences, (ii) the listener had a 'strong' prediction (e.g., a probability distribution with low entropy, where one continuation has much higher probability than all others) for the predictable sentence, but it did not match what the experimenter selected as predictable, or (iii) the listener had a 'weak' prediction (e.g., a probability distribution with higher entropy). The prediction for unpredictable sentences, in the latter two scenarios, might correspond to a (close to) uniform distribution over continuations, in line with a view that prediction is essential if not necessarily specific. By analogy to a neural network or statistical model which improves its predictions as it receives more input, learners with little experience – or experience with data from a different generative process – may engage in

prediction, but those predictions may deviate from what the experimenter assumes (as in scenarios ii and iii). The resulting eye-tracking and electroencephalography (EEG) patterns (assuming typical current approaches) would be indistinguishable from scenarios of absence (or diminished frequency) of prediction. It is worth noting that multivariate approaches provide one potential avenue for characterizing the distribution over predictions [37] and disentangling at least scenarios (ii) and (iii).

In terms of mechanism, the 'essential computation' view benefits from parsimony. On this view, continuously predicting and updating the internal model is the ubiquitous and obligatory dynamic of the language system, and any metabolic cost of prediction is subsumed within the cost of engaging in language comprehension and learning from experience. By contrast, 'optional' prediction accounts must – in addition to proposing a mechanism for prediction that is peripheral to the language comprehension/learning system – posit a mechanism by which the cost or utility of the prediction is tracked and evaluated. Executive function has often been proposed to fulfill that role (discussed in the following section and in Box 5).

Does linguistic prediction rely on language-specific or domain-general processes?—The term 'domain-general' is used to mean many different things in cognitive science – at least two of its uses are intersecting and relevant to understanding linguistic prediction at a mechanistic level. In one sense, prediction can be thought of as 'domain-general' because it is a 'canonical computation' of human cortex [121] that takes place in visual processing [122], sensorimotor learning [123], music [124], and social cognition [125], among many other domains. In another sense, however, it may be 'domain-specific' if it is implemented in local circuits rather than being directed by a shared prediction hub [126]. These need not be fully mutually exclusive. Unpredictable linguistic stimuli are known to evoke their largest responses in the language network – a set of frontal and temporal brain regions which respond most strongly and selectively during language comprehension [127,128]. Thus, the computation of linguistic prediction likely takes place, at least in large part, within the local circuits of this language network. Whether additional networks (or hubs) are recruited for linguistic prediction is an open question.

Another sense of 'domain-general' refers to the engagement of cognitive processes, such as executive function and working memory (Box 5), which are often associated with increased mental effort regardless of the computational substrate [129] and can be functionally localized to a frontoparietal network in the brain (known as the multiple-demand network [130]). Whether linguistic prediction engages amodal executive function and/or working memory is an area of active research that has the potential to substantially constrain the hypothesis space of neural architectures for language (reviewed in [106]). If linguistic prediction recruits amodal executive function, then an architecture in which prediction is implemented wholly in local language circuits becomes unlikely, but a variety of hybrid architectures could be compatible.

The observations of reduced prediction in some populations (e.g., children, older adults, and second-language learners) have been used to argue for the role of executive function (EF)/working memory (WM) in prediction, based on the assumption that young children and

older adults have reduced EF/WM [88,89] and that the EF resources of second-language learners are already taxed. Attempts to relate individual differences in EF/WM to linguistic prediction have yielded mixed results [131–133]. Alternatively, executive resources may come online to deal with the consequences of prediction rather than during generation. EEG studies have reported increases in power in theta band frequencies associated with unexpected words in a high-constraint context [134]. Activity in the theta band has been linked to both cognitive control and prediction error [135]. In sentence reading, power in the theta band may reflect the process of updating representations (*cf* [65]). Given the low spatial resolution of EEG, it is unknown whether these oscillatory effects reflect local computations within the language network or whether other systems are recruited.

Similarly, it appears that attention – sometimes considered to fall within the umbrella of executive resources, although it can be dissociated from inhibitory control or working memory [136] – appears to be needed to engage fully in predictive processing because listeners who are simultaneously distracted by a visual monitoring task show less facilitation from context in accessing the semantics of a word (i.e., smaller N400 effects [137]). Spectral analyses of EEG data have reported decreases in power in alpha/beta band frequencies before the target word when the context is constraining [134,138,139]. Activity in the alpha/beta band has been linked to increased attention across a variety of domains [140], suggesting that attention could be particularly important for the generation of predictions or that constraining sentences are more attention-grabbing.

By contrast, during naturalistic story listening, the relationship between neural activity and word-by-word surprisal is robust within the language system, but not within the multiple-demand system [128]. In addition, behavioral measures of comprehension difficulty (reading time slow downs) elicit robust blood oxygen-level dependent (BOLD) responses within the language system, but not within the multiple-demand system [128,141,142]. These converging findings indicate that frontoparietal executive resources are not recruited during typical language comprehension, at least to a degree that is detectable by fMRI.

This does not rule out the possibility that executive resources might be recruited in cases where the input is anomalous or is perceived to be an error. For instance, words that are predictable in context but appear to contain an easily corrected error (e.g., ‘He went to deposit his check at the p̲ank’) elicit a P600 ERP component [143,144] which is thought to be part of a larger family of P300 components that are involved in probabilistic error monitoring across domains and modalities [145]. Alternatively, cognitive control may be necessary for tracking changes in context on a longer timescale ([146,147]; *cf* [103]). The language network of the brain does not appear to be sensitive to linguistic context beyond (approximately) the sentence level [148], but humans flexibly adapt their predictions to the speaker or larger discourse environment, even in the absence of long-term exposure [47,49]. For instance, knowing where someone lives can immediately change our prediction of what they might say in a sentence such as ‘In the morning, I hear a lot of … [cars, birds, etc.]’, even if this is the first thing we have ever heard them say. This flexibility may rely on interactions between the language network and the domain-general multiple-demand network or further domain-specific networks such as the theory of mind network which is selectively engaged when humans reason about the mental states of others [149,150].

What is the relationship between prediction and memory?—What we predict is shaped by what we remember [151]. In the moment of processing, comprehenders derive predictions based on a representation of the context that is limited by the lossiness of human memory [152,153]. Over the course of development, children learn internal models that allow them to adequately use memory to predict the language they experience in their environment. However, the learning does not stop there (discussed in the section on Constraints and variability in prediction). Our mnemonic representations and predictions are continuously molded by what we read, watch, and discuss in everyday life. This is most evident when we consider how the predictions of 'experts' in a specialized domain differ from those of non-experts [154,155].

How does this learning take place? According to error-based learning accounts, every prediction instance is followed by a comparison between the prediction and the received input, and the discrepancy between the two is propagated through the network leading to an updated internal model. The updated model, in turn, determines both how a future context will be remembered and the prediction that a given context will generate. This class of accounts (in its simplest form) predicts that each instance of prediction also constitutes a learning event: when the prediction is far from the input, the updating is substantial, and when the prediction is close, it is minimal. Other adaptation mechanisms are also possible (e.g., homeostasis maintenance, reinforcement learning), and these should entail alternative testable hypotheses about the consequences of linguistic prediction.

One hypothesized neural signature of this update as it is being propagated through cortex is the late frontal positivity which is observed when the comprehender receives an input that is plausible but distinct from what they likely predicted (because the context points strongly to a different completion) [156–158]. This component has been linked to inhibitory control processes [159], and is potentially consistent with a role for cognitive control in prediction error signaling [135].

In addition to the immediate consequences indexed by these late frontal positivities, prediction also has longer-term consequences for the listener's representation of the language in memory. Words that are highly predictable in context (e.g., 'Alfonso has started biking to work instead of driving his car') elicit smaller repetition effects (measured as a reduction in N400 amplitude) when they are read again later (in a different sentence) relative to those which were originally less predictable from the context ('Jason tried to make space for others by moving his car') ([160]; *cf* [161]). A learning mechanism that relies on prediction error is consistent with these results because a larger prediction error (when 'car' is unpredictable) leads to a larger change to network connections – memory – and future estimates [105]. Further, words that would have constituted predictable sentence completions but were never presented (participants instead saw an unexpected ending) elicit smaller 'pseudo-repetition' effects when they are later presented [162] and more false alarm responses in a subsequent memory test than unrelated words [163], suggesting that there are lingering effects of generating the prediction on the individual's internal model that are not (fully) eliminated by the update process. Whether such memory effects fall out of future mechanistic accounts may provide a useful target for comparison.

What is the role of the production system in prediction during comprehension?—Early accounts of prediction during language comprehension proposed that these predictions are implemented in the same circuits that are used for language production [13,39,164]. This proposal is consistent with how prediction and production are implemented within a neural network model (e.g., [6]) and with evidence that production fluency is correlated with ERP effect magnitude at the level of the individual [165]. When participants are engaged in an articulatory suppression task (repeating the syllable /ta/ out loud) during comprehension, their N400 effects to articles preceding a predictable versus unpredictable noun were reduced [166], suggesting a causal role for the production system. Moreover, reading (silently) a highly predictable word during comprehension appears to confer a larger mnemonic benefit than an unpredictable word, akin to producing that word out loud rather than reading silently [167]. In addition, reading a predictive context out loud rather than silently appears to facilitate prediction [168]. The state of a word in memory appears to be affected both by its activation via a preceding context and by its activation via generation, suggesting shared or, minimally, tightly integrated representations for prediction during comprehension and for production.

However, a recent version of the prediction-by-production perspective proposes that prediction takes place via the production system only under particular conditions when the costs of production are low [115]. This account distinguishes the obligatory derivation of a speaker's intention via comprehension and covert imitation from the optional processes that generate predictions by running the intention through the production system. In this regard, addressing the role of the production system may inform, and be informed by, the question of whether prediction is essential or optional (discussed in the section on Is prediction essential or optional?).

This issue also bears on another important question in language research: what is the relationship between language comprehension and language production? The low-level input/output representations of production (i.e., motor movements of speech articulation, writing, signing) and comprehension (vision, audition, touch) are distinct. Neuropsychological work has long suggested that the divide between production and comprehension occurs relatively upstream: patients with aphasia often have largely preserved comprehension in the presence of production deficits or, more rarely, the reverse [169]. However, the two systems must share some representations because they both tap into the same rich world and linguistic knowledge. Early (and lifelong) language acquisition involves learning to produce linguistic sequences that follow the patterns of what one has heard from other speakers [170]. In addition, production, if it is to have the desired effect of communicating intended information, must entail considering how the listener will interpret what is said [171,172]. Indeed, neuroimaging evidence from healthy individuals suggests that there is substantial overlap in the brain regions engaged during high-level language production and comprehension [173,174]. Given the current spatial resolution of fMRI, these findings do not rule out the possibility that circuits for comprehension and production are separate but are connected and tightly interdigitated within the language network. Emerging techniques such as laminar fMRI and intracranial recordings promise to shed light on this key question.

Most mechanistic proposals focus on either comprehension or production (*cf* [13]). As the degree of interconnectedness of these systems is more thoroughly understood, mechanistic models optimized for both the goals of comprehension and production (e.g., saying the right words in the right sequence such that the listener will infer the intended meaning, while minimizing effort) may yield novel insights and constraints that have been missing from accounts of the role of prediction in comprehension alone.

Concluding remarks

In daily conversations, when reading the news, or engaging in most common forms of language comprehension, humans appear to concurrently predict. They use all cues available to them – preceding words, world knowledge, experience with the speaker, etc. – to put themselves in the right state for the input that will come next, and their predictions can encompass both the meaning and the form. Predictive processing of this type is thought to facilitate rapid exchanges of information, reduce (potentially costly) uncertainty, and enable lifelong learning. Despite compelling empirical evidence and a well-supported computational description, we do not yet know how the computations are constrained by the underlying mechanisms.

In the current review we have outlined key issues that, in our view, can help to guide the development of a mechanistic account of linguistic prediction. These include (i) determining whether prediction is essential and constitutive for language processing, as opposed to an optional strategy, (ii) delineating the role of non-linguistic mechanisms in linguistic prediction, (iii) explaining the bidirectional effects of memory on prediction, and of prediction on memory, and (iv) identifying the extent of interconnectedness between language production and language comprehension. Progress on these issues will circumscribe the space of (artificial and cortical) architectures and learning mechanisms that are considered.

Understanding linguistic prediction is a goal that lives at the heart of the cognitive sciences. Its pursuit draws on evidence and methods from psychology, linguistics, neuroscience, philosophy, and artificial intelligence. It has been, and continues to be, at the center of theoretical debates about the nature and development of human and machine intelligence. The quest to understand linguistic prediction at the mechanistic level promises to unify the different strands of research which have thus far operated in partial isolation, sometimes at different levels of analysis, and opens up a new set of questions at their intersection (see Outstanding questions for examples), suggesting that it will continue to have an equally profound influence on the field in the future.

Acknowledgments

This work was supported by grant NIA R15 AG073948 from the National Institutes of Health/National Institute on Aging to R.R. We are grateful to Robert Hawkins, Ina Bornkessel-Schlesewsky, Gina Kuperberg, and one anonymous reviewer for insightful comments which greatly improved the paper. We are also grateful to Michael Spivey and Cory Shain for illuminating discussions on the topic of prediction in language and for providing detailed comments on a previous draft of the manuscript.

References

1. Stivers T et al. (2009) Universals and cultural variation in turn-taking in conversation. Proc. Natl. Acad. Sci. U. S. A 106, 10587–10592 [PubMed: 19553212]
2. Brehm L and Meyer AS (2021) Planning when to say: dissociating cue use in utterance initiation using cross-validation. J. Exp. Psychol. Gen 150, 1772–1799 [PubMed: 33734778]
3. Elman JL (1990) Finding structure in time. Cogn. Sci 14, 179–211
4. Schrimpf M et al. (2021) The neural architecture of language: integrative modeling converges on predictive processing. Proc. Natl. Acad. Sci. U. S. A 118, e2105646118 [PubMed: 34737231]
5. Goldstein A et al. (2022) Shared computational principles for language processing in humans and deep language models. Nat. Neurosci 25, 369–380 [PubMed: 35260860]
6. Fitz H and Chang F (2019) Language ERPs reflect learning through prediction error propagation. Cogn. Psychol 111, 15–52 [PubMed: 30921626]
7. Clark A (2013) Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behav. Brain Sci 36, 181–204 [PubMed: 23663408]
8. de Lange FP et al. (2018) How do expectations shape perception? Trends Cogn. Sci 22, 764–779 [PubMed: 30122170]
9. Marr D and Poggio T (1976) From understanding computation to understanding neural circuitry. DSpace Published online October 1, 2004. https://dspace.mit.edu/handle/1721.1/5782
10. Rao RPN and Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci 2, 79–87 [PubMed: 10195184]
11. Falandays JB et al. (2021) Is prediction nothing more than multiscale pattern completion of the future? Brain Res. 1768, 147578 [PubMed: 34284021]
12. Luthra S et al. (2021) Does signal reduction imply predictive coding in models of spoken word recognition? Psychon. Bull. Rev 28, 1381–1389 [PubMed: 33852158]
13. Dell GS and Chang F (2014) The P-chain: relating sentence production and its disorders to comprehension and acquisition. Philos. Trans. R. Soc. B Biol. Sci 369, 20120394
14. Kuperberg GR and Jaeger TF (2016) What do we mean by prediction in language comprehension? Lang. Cogn. Neurosci 31, 32–59 [PubMed: 27135040]
15. Fischler I and Bloom PA (1979) Automatic and attentional processes in the effects of sentence contexts on word recognition. J. Verbal Learn. Verbal Behav 18, 1–20
16. Ehrlich SF and Rayner K (1981) Contextual effects on word perception and eye movements during reading. J. Verbal Learn. Verbal Behav 20, 641 −655
17. Kutas M and Hillyard SA (1984) Brain potentials during reading reflect word expectancy and semantic association. Nature 307, 161–163 [PubMed: 6690995]
18. Altmann GTM and Kamide Y (1999) Incremental interpretation at verbs: restricting the domain of subsequent reference. Cognition 73, 247–264 [PubMed: 10585516]
19. Tanenhaus MK et al. (1995) Integration of visual and linguistic information in spoken language comprehension. Science 268, 1632–1634 [PubMed: 7777863]
20. Smith NJ and Levy R (2013) The effect of word predictability on reading time is logarithmic. Cognition 128, 302–319 [PubMed: 23747651]
21. Aurnhammer C and Frank SL (2019) Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. Neuropsychologia 134, 107198 [PubMed: 31553896]
22. Frank SL et al. (2015) The ERP response to the amount of information conveyed by words in sentences. Brain Lang. 140, 1–11 [PubMed: 25461915]
23. Levy R (2008) Expectation-based syntactic comprehension. Cognition 106, 1126–1177 [PubMed: 17662975]
24. Hale J (2001) A probabilistic Earley parser as a psycholinguistic model. In NAACL ‘01: Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies. article 1073357, Association for Computing Machinery

25. Brothers T and Kuperberg GR (2021) Word predictability effects are linear, not logarithmic: implications for probabilistic models of sentence comprehension. J. Mem. Lang 116, 104174 [PubMed: 33100508]
26. Lowder MW et al. (2018) Lexical predictability during natural reading: effects of surprisal and entropy reduction. Cogn. Sci 42, 1166–1183 [PubMed: 29442360]
27. Szewczyk JM and Federmeier KD (2022) Context-based facilitation of semantic access follows both logarithmic and linear functions of stimulus probability. J. Mem. Lang 123,104311 [PubMed: 36337731]
28. Shain C et al. (2022) Large-scale evidence for logarithmic effects of word predictability on reading time. PsyArXiv Published online November, 24, 2022. 10.31234/osf.io/4hyna
29. Jain S and Huth A (2018) Incorporating context into language encoding models for fMRI. In Advances in Neural Information Processing Systems (NeurIPS 2018) (Bengio S et al., eds), pp. 6628–6637, NeurIPS
30. Federmeier KD and Kutas M (1999) A rose by any other name: long-term memory structure and sentence processing. J. Mem. Lang 41,469–495
31. Mantegna F et al. (2019) Distinguishing integration and prediction accounts of ERP N400 modulations in language processing through experimental design. Neuropsychologia 134,107199 [PubMed: 31545965]
32. Nieuwland M et al. (2019) Dissociable effects of prediction and integration during language comprehension: evidence from a large-scale study using brain potentials. Philos. Trans. R. Soc. B Biol. Sci 375, 20180522
33. Van Berkum JJA et al. (2005) Anticipating upcoming words in discourse: evidence from ERPs and reading times. J. Exp. Psychol. Learn. Mem. Cogn 31,443–467 [PubMed: 15910130]
34. Ferreira F and Chantavarin S (2018) Integration and prediction in language processing: a synthesis of old and new. Curr. Dir. Psychol. Sci 27, 443–448 [PubMed: 31130781]
35. Kriegeskorte N et al. (2008) Representational similarity analysis – connecting the branches of systems neuroscience. Front. Syst. Neurosci 2, 4 [PubMed: 19104670]
36. Wang L et al. (2018) Specific lexico-semantic predictions are associated with unique spatial and temporal patterns of neural activity. Elife 7, e39061 [PubMed: 30575521]
37. Wang L et al. (2020) Neural evidence for the prediction of animacy features during language comprehension: evidence from MEG and EEG representational similarity analysis. J. Neurosci 40, 3278–3291 [PubMed: 32161141]
38. Hubbard RJ and Federmeier KD (2021) Representational pattern similarity of electrical brain activity reveals rapid and specific prediction during language comprehension. Cereb. Cortex 31, 4300–4313 [PubMed: 33895819]
39. Federmeier KD (2007) Thinking ahead: the role and roots of prediction in language comprehension. Psychophysiology 44, 491–505 [PubMed: 17521377]
40. Lopopolo A et al. (2017) Using stochastic language models (SLM) to map lexical, syntactic, and phonological information processing in the brain. PLoS ONE 12, e0177794 [PubMed: 28542396]
41. Brennan JR and Hale JT (2019) Hierarchical structure guides rapid linguistic predictions during naturalistic listening. PLoS ONE 14, e0207741 [PubMed: 30650078]
42. Smith NJ and Levy R (2011) Cloze but no cigar: the complex relationship between cloze, corpus, and subjective probabilities in language processing. Proc. Cogn. Sci. Soc 33, 1637–1642
43. Broderick MP et al. (2018) Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. Curr. Biol 28, 803–809 [PubMed: 29478856]
44. Frank SL and Willems RM (2017) Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. Lang. Cogn. Neurosci 32, 1192–1203
45. Michaelov JA et al. (2023) Strong prediction: language model surprisal explains multiple N400 effects. Neurobiol. Lang 5, 1–29. 10.1162/nol_a_00105
46. Hagoort P et al. (2004) Integration of word meaning and world knowledge in language comprehension. Science 304, 438–441 [PubMed: 15031438]
47. Nieuwland M and Van Berkum JJA (2006) When peanuts fall in love: N400 evidence for the power of discourse. J. Cogn. Neurosci 18, 1098–1111 [PubMed: 16839284]

48. Lowder MW and Ferreira F (2019) I see what you meant to say: anticipating speech errors during online sentence processing. J. Exp. Psychol. Gen 148, 1849–1858 [PubMed: 30556724]
49. Ryskin R et al. (2020) Talker-specific predictions during language processing. Lang. Cogn. Neurosci 35, 797–812 [PubMed: 33693050]
50. Bosker HR et al. (2019) Counting 'uhm's: how tracking the distribution of native and non-native disfluencies influences online language comprehension. J. Mem. Lang 106, 189–202
51. Hanulíková A et al. (2012) When one person's mistake is another's standard usage: the effect of foreign accent on syntactic processing. J. Cogn. Neurosci 24, 878–887 [PubMed: 21812565]
52. Brothers T et al. (2019) Flexible predictions during listening comprehension: speaker reliability affects anticipatory processes. Neuropsychologia 135,107225 [PubMed: 31605686]
53. Delaney-Busch N et al. (2019) Neural evidence for Bayesian trial-by-trial adaptation on the N400 during semantic priming. Cognition 187, 10–20 [PubMed: 30797099]
54. Ness T and Meltzer-Asscher A (2021) Rational adaptation in lexical prediction: the influence of prediction strength. Front. Psychol 12, 622873 [PubMed: 33935874]
55. Nieuwland M (2020) How 'rational' is semantic prediction? A discussion and reanalysis of Delaney-Busch, Morgan, Lau & Kuperberg (2019). PsyArXiv Published online March, 03, 2020. 10.31234/osf.io/sm7fq
56. Nieuwland M (2021)Commentary: rational adaptation in lexical prediction: the influence of prediction strength. Front. Psychol 12, 735849 [PubMed: 34504467]
57. van Wonderen E and Nieuwland MS (2023) Lexical prediction does not rationally adapt to prediction error: ERP evidence from pre-nominal articles. J. Mem. Lang 132, 104435
58. Federmeier KD (2022) Connecting and considering: electrophysiology provides insights into comprehension. Psychophysiology 59, e13940 [PubMed: 34520568]
59. Frisson S et al. (2017) No prediction error cost in reading: evidence from eye movements. J. Mem. Lang 95, 200–214
60. Caucheteux C et al. (2023) Evidence of a predictive coding hierarchy in the human brain listening to speech. Nat. Hum. Behav 7, 430–441 [PubMed: 36864133]
61. Heilbron M et al. (2022) A hierarchy of linguistic predictions during natural language comprehension. Proc. Natl. Acad. Sci. U. S. A 119, e2201968119 [PubMed: 35921434]
62. Schmitt L-M et al. (2021) Predicting speech from a cortical hierarchy of event-based time scales. Sci. Adv 7, eabi6070 [PubMed: 34860554]
63. Weissbart H et al. (2020) Cortical tracking of surprisal during continuous speech comprehension. J. Cogn. Neurosci 32, 155–166 [PubMed: 31479349]
64. Broderick MP et al. (2019) Semantic context enhances the early auditory encoding of natural speech. J. Neurosci 39, 7564–7575 [PubMed: 31371424]
65. Donhauser PW and Baillet S (2019) Two distinct neural timescales for predictive speech processing. Neuron 105, 385–393 [PubMed: 31806493]
66. Armeni K et al. (2019) Frequency-specific brain dynamics related to prediction during language comprehension. Neuroimage 198, 283–295 [PubMed: 31100432]
67. DeLong KA et al. (2005) Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. Nat. Neurosci 8,1117–1121 [PubMed: 16007080]
68. Nicenboim B et al. (2019) Are words pre-activated probabilistically during sentence comprehension? Evidence from new data and a Bayesian random-effects meta-analysis using publicly available data. Neuropsychologia 142,107427
69. Nieuwland M et al. (2018) Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. Elife 7, e33468 [PubMed: 29631695]
70. Nieuwland M et al. (2020) Anticipating words during spoken discourse comprehension: a large-scale, pre-registered replication study using brain potentials. Cortex 133, 1–36 [PubMed: 33096395]
71. Yan S et al. (2017) Prediction (or not) during language processing. A commentary on Nieuwland et al. (2017) and DeLong et al. (2005). BioRxiv Published online May, 03, 2017. 10.1101/143750
72. Rabovsky M (2020) Change in a probabilistic representation of meaning can account for N400 effects on articles: a neural network model. Neuropsychologia 143, 107466 [PubMed: 32315697]

73. Szewczyk JM et al. (2020) The mechanisms of prediction updating that impact the processing of upcoming word: an event-related potential study on sentence comprehension. J. Exp. Psychol. Learn. Mem. Cogn 46, 1714–1734 [PubMed: 32297790]
74. Fleur DS et al. (2020) Definitely saw it coming? The dual nature of the pre-nominal prediction effect. Cognition 204,104335 [PubMed: 32619896]
75. Nieuwland M (2019) Do 'early' brain responses reveal word form prediction during language comprehension? A critical review. Neurosci. Biobehav. Rev 96, 367–400 [PubMed: 30621862]
76. Knill DC and Pouget A (2004) The Bayesian brain: the role of uncertainty in neural coding and computation. Trends Neurosci. 27, 712–719 [PubMed: 15541511]
77. Kwisthout J and van Rooij I (2020) Computational resource demands of a predictive Bayesian brain. Comput. Brain Behav 3, 174–188
78. Lieder F and Griffiths TL (2020) Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. Behav. Brain Sci 43, e1
79. Bornkessel-Schlesewsky I et al. (2022) Rapid adaptation of predictive models during language comprehension: aperiodic EEG slope, individual alpha frequency and idea density modulate individual differences in real-time model updating. Front. Psychol 13, 817516 [PubMed: 36092106]
80. Elman JL (2009) On the meaning of words and dinosaur bones: lexical knowledge without a lexicon. Cogn. Sci 33, 547–582 [PubMed: 19662108]
81. Ylinen S et al. (2017) Predictive coding accelerates word recognition and learning in the early stages of language development. Dev. Sci 20, e12472
82. Reuter T et al. (2019) Predict and redirect: prediction errors support children's word learning. Dev. Psychol 55,1656–1665 [PubMed: 31094555]
83. Gambi C et al. (2021) The relation between preschoolers' vocabulary development and their ability to predict and recognize words. Child Dev. 92, 1048–1066 [PubMed: 32865231]
84. Borovsky A et al. (2012) Knowing a lot for one's age: vocabulary skill and not age is associated with anticipatory incremental sentence interpretation in children and adults. J. Exp. Child Psychol 112, 417–436 [PubMed: 22632758]
85. Gambi C et al. (2018) The development of linguistic prediction: predictions of sound and meaning in 2- to 5-year-olds. J. Exp. Child Psychol 173, 351 −370 [PubMed: 29793772]
86. Reuter T et al. (2018) Individual differences in nonverbal prediction and vocabulary size in infancy. Cognition 176, 215–219 [PubMed: 29604470]
87. Rabagliati H et al. (2016) Learning to predict or predicting to learn? Lang. Cogn. Neurosci 31,94–105
88. Davidson MC et al. (2006) Development of cognitive control and executive functions from 4 to 13 years: evidence from manipulations of memory, inhibition, and task switching. Neuropsychologia 44, 2037–2078 [PubMed: 16580701]
89. Hartshorne JK and Germine LT (2015) When does cognitive functioning peak? The asynchronous rise and fall of different cognitive abilities across the life span. Psychol. Sci 26, 433–443 [PubMed: 25770099]
90. Reuter T et al. (2022) Look at that: spatial deixis reveals experience-related differences in prediction. Lang. Acquis 29, 1 –26 [PubMed: 35281590]
91. Ito A et al. (2017) On predicting form and meaning in a second language. J. Exp. Psychol. Learn. Mem. Cogn 43, 635–652 [PubMed: 27668483]
92. Ito A et al. (2018) Investigating the time-course of phonological prediction in native and non-native speakers of English: a visual world eye-tracking study. J. Mem. Lang 98, 1–11
93. Ng S et al. (2017) Use of contextual information and prediction by struggling adult readers: evidence from reading times and event-related potentials. Sci. Stud. Read 21,359–375
94. Peters RE et al. (2018) Vocabulary size and native speaker self-identification influence flexibility in linguistic prediction among adult bilinguals. Appl. Psycholinguist 39, 1439–1469 [PubMed: 31105359]
95. Dijkgraaf A et al. (2017) Predicting upcoming information in native-language and non-native-language auditory word recognition. Biling. Lang. Cogn 20, 917–930

96. Ito A et al. (2018) A cognitive load delays predictive eye movements similarly during L1 and L2 comprehension. Biling. Lang. Cogn 21,251–264

97. Chun E and Kaan E (2019) L2 prediction during complex sentence processing. J. Cult. Cogn. Sci 3, 203–216

98. Payne BR and Federmeier KD (2018) Contextual constraints on lexico-semantic processing in aging: evidence from single-word event-related brain potentials. Brain Res. 1687, 117–128 [PubMed: 29462609]

99. Cheimariou S et al. (2019) Lexical prediction in the aging brain: the effects of predictiveness and congruency on the N400 ERP component. Neuropsychol. Dev. Cogn. B Aging Neuropsychol. Cogn 26, 781 –806 [PubMed: 30293520]

100. Federmeier KD and Kutas M (2019) What's 'left'? Hemispheric sensitivity to predictability and congruity during sentence reading by older adults. Neuropsychologia 133, 107173 [PubMed: 31430444]

101. Broderick MP et al. (2021) Dissociable electrophysiological measures of natural language processing reveal differences in speech comprehension strategy in healthy ageing. Sci. Rep 11, 4963 [PubMed: 33654202]

102. Choi W et al. (2017) Effects of word predictability and preview lexicality on eye movements during reading: a comparison between young and older adults. Psychol. Aging 32, 232–242 [PubMed: 28333501]

103. Dave S et al. (2018) Electrophysiological evidence for preserved primacy of lexical prediction in aging. Neuropsychologia 117, 135–147 [PubMed: 29852201]

104. Huettig F and Janse E (2016) Individual differences in working memory and processing speed predict anticipatory spoken language processing in the visual world. Lang. Cogn. Neurosci 31,80–93

105. Rabovsky M et al. (2018) Modelling the N400 brain potential as change in a probabilistic representation of meaning. Nat. Hum. Behav 2, 693 [PubMed: 31346278]

106. Ryskin R et al. (2020) Do domain-general executive resources play a role in linguistic prediction? Re-evaluation of the evidence and a path forward. Neuropsychologia 136, 107258 [PubMed: 31730774]

107. Moran RJ et al. (2014) The brain ages optimally to model its environment: evidence from sensory learning over the adult lifespan. PLoS Comput Biol 10, e1003422 [PubMed: 24465195]

108. Verhaeghen P (2003) Aging and vocabulary score: a metaanalysis. Psychol. Aging 18, 332–339 [PubMed: 12825780]

109. Wang L et al. (2023) Predictive coding across the left frontotemporal hierarchy during language comprehension. Cereb. Cortex 33, 4478–4497 [PubMed: 36130089]

110. Bornkessel-Schlesewsky I and Schlesewsky M (2019) Toward a neurobiologically plausible model of language-related, negative event-related potentials. Front. Psychol 10, 298 [PubMed: 30846950]

111. Nour Eddine S et al. (2022) The N400 in silico: a review of computational models. Psychol. Learn. Motiv 76, 123–206

112. Rabovsky M and McRae K (2014) Simulating the N400 ERP component as semantic network error: Insights from a feature-based connectionist attractor model of word meaning. Cognition 132, 68–89 [PubMed: 24762924]

113. McClelland JL and Elman JL (1986) The TRACE model of speech perception. Cogn. Psychol 18, 1–86 [PubMed: 3753912]

114. Lupyan G and Clark A (2015) Words and the world: predictive coding and the language-perception-cognition interface. Curr. Dir. Psychol. Sci 24, 279–284

115. Pickering MJ and Gambi C (2018) Predicting while comprehending language: a theory and review. Psychol. Bull 144, 1002–1044 [PubMed: 29952584]

116. Huettig F and Pickering MJ (2019) Literacy advantages beyond reading: prediction of spoken language. Trends Cogn. Sci 23, 464–475 [PubMed: 31097411]

117. Grisoni L et al. (2017) Neural correlates of semantic prediction and resolution in sentence processing. J. Neurosci 37, 4848–4858 [PubMed: 28411271]

118. Freunberger D and Roehm D (2017) The costs of being certain: brain potential evidence for linguistic preactivation in sentence processing. Psychophysiology 54, 824–832 [PubMed: 28240780]
119. Huettig F and Guerra E (2019) Effects of speech rate, preview time of visual context, and participant instructions reveal strong limits on prediction in language processing. Brain Res. 1706, 196–208 [PubMed: 30439351]
120. Brothers T et al. (2017) Goals and strategies influence lexical prediction during sentence comprehension. J. Mem. Lang 93, 203–216
121. Keller GB and Mrsic-Flogel TD (2018) Predictive processing: a canonical cortical computation. Neuron 100, 424–435 [PubMed: 30359606]
122. Kok P et al. (2017) Prior expectations induce prestimulus sensory templates. Proc. Natl. Acad. Sci. U. S. A 114, 10473–10478 [PubMed: 28900010]
123. Meirhaeghe N et al. (2021) A precise and adaptive neural mechanism for predictive temporal processing in the frontal cortex. Neuron 109, 2995–3011 [PubMed: 34534456]
124. Patel AD and Morgan E (2017) Exploring cognitive relations between prediction in language and music. Cogn. Sci 41, 303–320 [PubMed: 27665745]
125. Koster-Hale J and Saxe R (2013) Theory of mind: a neural prediction problem. Neuron 79, 836–848 [PubMed: 24012000]
126. Fedorenko E and Shain C (2021) Similarity of computations across domains does not imply shared implementation: the case of language comprehension. Curr. Dir. Psychol. Sci 30, 526–534 [PubMed: 35295820]
127. Fedorenko E et al. (2011) Functional specificity for high-level linguistic processing in the human brain. Proc. Natl. Acad. Sci. U. S. A 108, 16428–16433 [PubMed: 21885736]
128. Shain C et al. (2020) fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. Neuropsychologia 138, 107307 [PubMed: 31874149]
129. Duncan J (2010) The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. Trends Cogn. Sci. 14,172–179 [PubMed: 20171926]
130. Fedorenko E et al. (2013) Broad domain generality in focal regions of frontal and parietal cortex. Proc. Natl. Acad. Sci. U. S. A 110, 16616–16621 [PubMed: 24062451]
131. James AN et al. (2018) Individual differences in syntactic processing: Is there evidence for reader-text interactions? J. Mem. Lang 102, 155–181 [PubMed: 30713367]
132. Ness T and Meltzer-Asscher A (2018) Predictive preupdating and working memory capacity: evidence from event-related potentials. J. Cogn. Neurosci 30, 1916–1938 [PubMed: 30125220]
133. Zirnstein M et al. (2018) Cognitive control ability mediates prediction costs in monolinguals and bilinguals. Cognition 176, 87–106 [PubMed: 29549762]
134. Rommers J et al. (2017) Alpha and theta band dynamics related to sentential constraint and word expectancy. Lang. Cogn. Neurosci 32, 576–589 [PubMed: 28761896]
135. Cavanagh JF and Frank MJ (2014) Frontal theta as a mechanism for cognitive control. Trends Cogn. Sci 18, 414–421 [PubMed: 24835663]
136. Panichello MF and Buschman TJ (2021) Shared mechanisms underlie the control of working memory and attention. Nature 592, 601 –605 [PubMed: 33790467]
137. Hubbard RJ and Federmeier KD (2021) Dividing attention influences contextual facilitation and revision during language comprehension. Brain Res. 1764,147466 [PubMed: 33861998]
138. Wang L et al. (2018) Language prediction is reflected by coupling between frontal gamma and posterior alpha oscillations. J. Cogn. Neurosci. 30, 432–447 [PubMed: 28949823]
139. León-Cabrera P et al. (2022) Alpha power decreases associated with prediction in written and spoken sentence comprehension. Neuropsychologia 173,108286 [PubMed: 35679987]
140. Haegens S et al. (2012) Somatosensory anticipatory alpha activity increases to suppress distracting input. J. Cogn. Neurosci 24, 677–685 [PubMed: 22066587]
141. Shain C et al. (2022) Robust effects of working memory demand during naturalistic language comprehension in language-selective cortex. J. Neurosci 42, 7412–7430 [PubMed: 36002263]

142. Wehbe L et al. (2021) Incremental language comprehension difficulty predicts activity in the language network but not the multiple demand network. Cereb. Cortex 31,4006–4023 [PubMed: 33895807]
143. Laszlo S and Federmeier KD (2009) A beautiful day in the neighborhood: an event-related potential study of lexical relationships and prediction in context. J. Mem. Lang. 61,326–338 [PubMed: 20161064]
144. Ryskin R et al. (2021) An ERP index of real-time error correction within a noisy-channel framework of human communication. Neuropsychologia 158, 107855 [PubMed: 33865848]
145. Leckey M and Federmeier KD (2019) The P3b and P600(s): positive contributions to language comprehension. Psychophysiology 57, e13351 [PubMed: 30802979]
146. Miller EK and Cohen JD (2001)An integrative theory of prefrontal cortex function. Annu. Rev. Neurosci 24, 167–202 [PubMed: 11283309]
147. Noelle DC (2012) On the neural basis of rule-guided behavior. J. Integr. Neurosci 11, 453–475 [PubMed: 23351052]
148. Blank IA and Fedorenko E (2020) No evidence for differences among language regions in their temporal receptive windows. NeuroImage 219, 116925 [PubMed: 32407994]
149. Paunov AM et al. (2019) Functionally distinct language and theory of mind networks are synchronized at rest and during language comprehension. J. Neurophysiol 121, 1244–1265 [PubMed: 30601693]
150. Saxe R and Kanwisher N (2003) People thinking about thinking people. The role of the temporo-parietal junction in 'theory of mind. NeuroImage 19, 1835–1842 [PubMed: 12948738]
151. Christiansen MH and Chater N (2016) The now-or-never bottleneck: a fundamental constraint on language. Behav. Brain Sci 39, e62 [PubMed: 25869618]
152. Futrell R et al. (2020) Lossy-context surprisal: an information-theoretic model of memory effects in sentence processing. Cogn. Sci 44, e12814 [PubMed: 32100918]
153. Hahn M et al. (2022) A resource-rational model of human processing of recursive linguistic structure. Proc. Natl. Acad. Sci. U. S. A 119, e2122602119 [PubMed: 36260742]
154. Troyer M et al. (2020) Lumos!: electrophysiological tracking of (wizarding) world knowledge use during reading. J. Exp. Psychol. Learn. Mem. Cogn 46, 476–486 [PubMed: 31294584]
155. Troyer M and Kutas M (2020) To catch a snitch: brain potentials reveal variability in the functional organization of (fictional) world knowledge during reading. J. Mem. Lang 113, 104111 [PubMed: 33678947]
156. DeLong KA and Kutas M (2020) Comprehending surprising sentences: sensitivity of post-N400 positivities to contextual congruity and semantic relatedness. Lang. Cogn. Neurosci 35, 1044–1063 [PubMed: 36176318]
157. Federmeier KD et al. (2007) Multiple effects of sentential constraint on word processing. Brain Res 1146, 75–84 [PubMed: 16901469]
158. Van Petten C and Luka BJ (2012) Prediction during language comprehension: benefits, costs, and ERP components. Int J. Psychophysiol. 83, 176–190 [PubMed: 22019481]
159. Ness T and Meltzer-Asscher A (2018) Lexical inhibition due to failed prediction: behavioral evidence and ERP correlates. J. Exp. Psychol. Learn. Mem. Cogn 44, 1269–1285 [PubMed: 29283606]
160. Rommers J and Federmeier KD (2018) Predictability's aftermath: downstream consequences of word predictability as revealed by repetition effects. Cortex 101,16–30 [PubMed: 29414458]
161. Lai MK et al. (2021)The fate of the unexpected: consequences of misprediction assessed using ERP repetition effects. Brain Res. 1757, 147290 [PubMed: 33516812]
162. Rommers J and Federmeier KD (2018) Lingering expectations: a pseudo-repetition effect for words previously expected but not presented. NeuroImage 183, 263–272 [PubMed: 30107258]
163. Hubbard RJ et al. (2019) Downstream behavioral and electro-physiological consequences of word prediction on recognition memory. Front. Hum. Neurosci 13, 291 [PubMed: 31555111]
164. Pickering MJ and Garrod S (2013) An integrated theory of language production and comprehension. Behav. Brain Sci 36, 329–347 [PubMed: 23789620]

165. Federmeier KD et al. (2010) Age-related and individual differences in the use of prediction during language comprehension. Brain Lang. 115, 149–161 [PubMed: 20728207]
166. Martin CD et al. (2018) Prediction is production: the missing link between language production and comprehension. Sci. Rep 8, 1079 [PubMed: 29348611]
167. Rommers J et al. (2020) Word predictability blurs the lines between production and comprehension: evidence from the production effect in memory. Cognition 198,104206 [PubMed: 32035323]
168. Lelonkiewicz JR et al. (2021) EXPRESS: the role of language production in making predictions during comprehension. Q. J. Exp. Psychol 74, 2193–2209
169. Kertesz A and Poole E (1974) The aphasia quotient: the taxonomic approach to measurement of aphasic disability. Can. J. Neurol. Sci 1,7–16 [PubMed: 4434266]
170. MacDonald MC (2013) How language production shapes language form and comprehension. Front. Psychol 4, 226 [PubMed: 23637689]
171. Brown-Schmidt S and Heller D (2018) Perspective-taking during conversation. In The Oxford Handbook of Psycholinguistics (Rueschemeyer S-A and Gaskell MG, eds), pp. 548–572, Oxford University Press
172. Frank MC and Goodman ND (2012) Predicting pragmatic reasoning in language games. Science 336, 998 [PubMed: 22628647]
173. Hu J et al. (2022) Precision fMRI reveals that the language-selective network supports both phrase-structure building and lexical access during language production. Cereb. Cortex 33, 4384–4404
174. Silbert LJ et al. (2014) Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. Proc. Natl. Acad. Sci. U. S. A 111, E4687–E4696 [PubMed: 25267658]
175. Oliver BM (1952) Efficient coding. Bell Syst Tech. J 31, 724–750
176. Barlow HB (1961) Possible principles underlying the transformations of sensory messages. In Sensory Communication (Rosenblith WA, ed.), pp. 216–234, MIT Press
177. Mumford D (1992) On the computational architecture of the neocortex. Biol. Cybern 66, 241–;251 [PubMed: 1540675]
178. Srinivasan MV et al. (1982) Predictive coding: a fresh view of inhibition in the retina. Proc. R. Soc. Lond. B Biol. Sci 216, 427–459 [PubMed: 6129637]
179. Davis MH and Sohoglu E (2020) Three functions of prediction error for Bayesian inference in speech perception. In The Cognitive Neurosciences (6th edn) (Gazzaniga M et al., eds), pp. 177–189, MIT Press
180. Friston K (2005) A theory of cortical responses. Philos. Trans. R. Soc. B Biol. Sci 360, 815–836
181. Ma WJ et al. (2006) Bayesian inference with probabilistic population codes. Nat. Neurosci 9, 1432–1438 [PubMed: 17057707]
182. Bastos AM et al. (2012) Canonical microcircuits for predictive coding. Neuron 76, 695–711 [PubMed: 23177956]
183. Chao ZC et al. (2018) Large-scale cortical networks for hierarchical prediction and prediction error in the primate brain. Neuron 100, 1252–1266 [PubMed: 30482692]
184. Dürschmid S et al. (2016) Hierarchy of prediction errors for auditory events in human temporal and frontal cortex. Proc. Natl. Acad. Sci. U. S. A 113, 6755–6760 [PubMed: 27247381]
185. Dürschmid S et al. (2019) Direct evidence for prediction signals in frontal cortex independent of prediction error. Cereb. Cortex 29, 4530–4538 [PubMed: 30590422]
186. Walsh KS et al. (2020) Evaluating the neurophysiological evidence for predictive processing as a model of perception. Ann. N. Y. Acad. Sci 1464, 242–268 [PubMed: 32147856]
187. Ali A et al. (2022) Predictive coding is a consequence of energy efficiency in recurrent neural networks. Patterns 3, 100639 [PubMed: 36569556]
188. Solomon SS et al. (2021)Limited evidence for sensory prediction error responses in visual cortex of macaques and humans. Cereb. Cortex 31,3136–3152 [PubMed: 33683317]
189. Fedorenko E et al. (2010) New method for fMRI investigations of language: defining ROIs functionally in individual subjects. J. Neurophysiol 104, 1177–1194 [PubMed: 20410363]

190. McMurray B (2023) I'm not sure that curve means what you think it means: toward a [more] realistic understanding of the role of eye-movement generation in the visual world paradigm. Psychon. Bull. Rev 30, 102–146 [PubMed: 35962241]
191. Taylor WL (1953) 'Cloze procedure':a new tool for measuring readability. J. Bull 30, 415–433
192. Caucheteux C and King J-R (2022) Brains and algorithms partially converge in natural language processing. Commun. Biol 5, 134 [PubMed: 35173264]
193. Antonello R and Huth A (2023) Predictive coding or just feature discovery? An alternative account of why language models fit brain data. Neurobiol. Lang 5, 00087
194. Guest O and Martin AE (2023) On logical inference over brains, behaviour, and artificial neural networks. Comput. Brain Behav 6, 213–227
195. Hale JT et al. (2022) Neurocomputational models of language processing. Annu. Rev. Linguist 8, 427–446
196. Tuckute G et al. (2023) Driving and suppressing the human language network using large language models. BioRxiv Published online August, 01, 2023. 10.1101/2023.04.16.537080
197. Lee CS et al. (2021) Anticipation of temporally structured events in the brain. Elife 10, e64972 [PubMed: 33884953]
198. Hayden BY et al. (2011) Surprise signals in anterior cingulate cortex: neuronal encoding of unsigned reward prediction errors driving adjustment in behavior. J. Neurosci 31,4178–4187 [PubMed: 21411658]
199. Carter CS et al. (1998) Anterior cingulate cortex, error detection, and the online monitoring of performance. Science 280,747–749 [PubMed: 9563953]
200. Just MA and Carpenter PA (1992) A capacity theory of comprehension: individual differences in working memory. Psychol. Rev 99, 122–149 [PubMed: 1546114]
201. Botvinick MM (2007) Multilevel structure in behaviour and in the brain: a model of Fuster's hierarchy. Philos. Trans. R. Soc. B Biol. Sci 362, 1615–1626
202. Kriete T et al. (2013) Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. Proc. Natl. Acad. Sci. U. S. A 110, 16390–16395 [PubMed: 24062434]
203. Friedman NP and Miyake A (2017) Unity and diversity of executive functions: individual differences as a window on cognitive structure. Cortex 86, 186–204 [PubMed: 27251123]
204. Gratton G et al. (2018) Dynamics of cognitive control: theoretical bases, paradigms, and a view for the future. Psychophysiology 55, e13016
205. Vincent JL et al. (2008) Evidence for a frontoparietal control system revealed by intrinsic functional connectivity. J. Neurophysiol 100, 3328–3342 [PubMed: 18799601]
206. Michalka SW et al. (2015) Short-term memory for space and time flexibly recruit complementary sensory-biased frontal lobe attention networks. Neuron 87, 882–892 [PubMed: 26291168]
207. Noyce AL et al. (2017) Sensory-biased and multiple-demand processing in human lateral frontal cortex. J. Neurosci 37, 8755–8766 [PubMed: 28821668]
208. Engelhardt PE et al. (2017) Executive function and intelligence in the resolution of temporary syntactic ambiguity: an individual differences investigation. Q. J. Exp. Psychol 70, 1263–1281
209. Farmer TA et al. (2017) Reading span task performance, linguistic experience, and the processing of unexpected syntactic events. Q. J. Exp. Psychol 70, 413–433
210. Van Dyke JA et al. (2014) Low working memory capacity is only spuriously related to poor reading comprehension. Cognition 131, 373–403 [PubMed: 24657820]
211. Blank I and Fedorenko E (2017) Domain-general brain regions do not track linguistic input as closely as language-selective regions. J. Neurosci 37, 9999–10011 [PubMed: 28871034]
212. Diachek E et al. (2020) The domain-general multiple demand (MD) network does not support core aspects of language comprehension: a large-scale fMRI investigation. J. Neurosci 40, 4536–4550 [PubMed: 32317387]
213. Nozari N (2018) How special is language production? Perspectives from monitoring and control. Psychol. Learn. Motiv 68, 179–213
214. Cohen JD and Servan-Schreiber D (1992) Context, cortex, and dopamine: a connectionist approach to behavior and biology in schizophrenia. Psychol. Rev 99, 45–77 [PubMed: 1546118]

Highlights

During language processing, comprehenders predict upcoming linguistic input. These predictions draw on many sources of information including the preceding sentence context.

We review a substantial body of evidence which has explored the information used and generated in the computation of prediction, as well as the constraints.

Further progress in understanding prediction in language comprehension will likely require developing mechanistic explanations. We discuss four research questions which may help to guide the development of these explanations.

Box 1.

Predictive coding

The idea that human minds and brains predict and learn from error has a long history in cognitive science. Inspired by information-theoretic approaches to efficient message/video transmission (e.g., [175]), Barlow [176] argued that perception encodes sensory stimuli in a way that reduces redundancy and transmits those signals which are least redundant, in other words those that are least predictable. These ideas were the direct precursors of the modern-day predictive coding account.

The term 'predictive coding' is sometimes used interchangeably with 'prediction' or 'predictive processing' to refer to any instance of perception or cognition where a response is reduced for input that is likely given 'top-down' information (e.g., the statistics of the environment) than input that is unlikely. We reserve the term 'predictive coding' here for the more specific instantiation of this framework proposed by Rao and Ballard [10] to account for the suppression of activity in some visual cortical neurons (*cf* /177,178]) that has been taken up by many others across domains of perception, action, and cognition ([8,125,179,180]; comprehensively reviewed in [7]). The key features of this proposal (summarized in Figure 1) are listed below.

- **i.** Hierarchical organization: sensory inputs (e.g., sound) are represented at the lowest levels, whereas more abstract information (e.g., word meanings) is represented at higher levels of the hierarchy (in simulations, these levels correspond to layers of a neural network model, but conceptually these levels could map onto different brain regions and/or different cortical layers within a region).
- **ii.** Top-down predictions are transmitted from higher levels to lower levels of the hierarchy: each level predicts the responses in the level immediately below via feedback connections.
- **iii.** Bottom-up prediction errors travel from lower levels to higher levels of the hierarchy: each level transmits the discrepancy between the predicted response and the estimated actual response to the level immediately above it via feedforward connections.
- **iv.** Prediction errors are used to update the response at each level and generate the next prediction, thus allowing the mind to process/infer the current input and continuously adapt to its environment.
- **v.** Predictions and errors are carried by distinct populations of functional units: the mapping to neuroanatomy is purely speculative at this point, but some functional distinction will be necessary to allow both forward cascades of errors and backward flow of predictions.
- **vi.** Bayesian inference and precision-weighting: the predictive coding algorithm approximates Bayesian inference. The outcome of predictive computations corresponds to probability distributions over predictions, as opposed to singular predictions, potentially in the form of probabilistic population

codes [181]. Following Bayesian principles, the posterior probabilities of the internal model, at each level, capture the uncertainty of both the bottom-up input and the top-down predictions. When the sensory input is noisy or prior knowledge is sparse, they are said to have low precision (defined as inverse variance) and are down-weighted in the update computation.

This proposal finds support in neuroanatomy [182], computational simulations which match response patterns of recordings directly from animal cortical cells (e.g., [10,183]), and a plethora of findings that human neural response patterns are consistent with predictive coding ([122,184,185]; comprehensively reviewed in [186]). Furthermore, recent computational work shows that a recurrent neural network trained to optimize energy efficiency, within a predictive environment, self-organizes into distinct subsets of neurons, one carrying errors and the other carrying predictions [187]. However, many of the empirical studies that are attributed to predictive coding (especially in higher-level cognitive domains) are also consistent with other possible models of predictive processing. Some empirical results are inconsistent with the predictive coding account in its current form [188].

Box 2.

Investigating prediction in language by measuring human neural activity

Event-related potentials (ERPs)/electroencephalography (EEG)/ magnetoencephalography (MEG)

Electrical activity at the scalp is recorded while participants read or listen to language. The EEG signal is time-locked to the onset ('event-related') of particularly crucial stimuli to compare neural responses, for instance, to words that are predictable or unpredictable in a sentence [17]. Unpredictable words typically elicit a more negative potential that peaks ~400 ms after it is presented – the N400 – relative to predictable words. Predictability also has downstream consequences for later components of the EEG, and the specific aspects of prediction/processing that are reflected in each component constitute an area of active investigation.

The high temporal precision of EEG allows neural responses to be measured as soon as words appear. On the other hand, EEG has very low spatial resolution, and inferences about the neural sources of ERPs are therefore coarse-grained at best. Determining the nature of what was predicted based on the ERP signal is challenging, and requires complex designs and/or multivariate analysis approaches. MEG is often used in analogous ways (the signal that is measured derives from magnetic fields rather than from electrical currents) but its spatial precision is improved relative to EEG.

Time-frequency analyses

EEG/MEG can also be decomposed into frequency bands – any complex waveform can be seen as a mixture of waves of differing frequencies. Bands of different frequency may encode/transmit different types of information, and these spectral analyses have the potential to reveal multiple ongoing mechanisms at work during language comprehension that cannot be detected by ERPs. For instance, theta band frequencies (4–7 Hz) have been proposed to carry prediction error signals and have been linked to the engagement of cognitive control [134].

Electrocorticography (ECoG)/intracranial recordings

Recent technological and neurosurgical advances allow electrical activity to be recorded directly from human cortex. Paradigms from EEG/MEG studies can be readily used with these methods, although this is an area of rapid development (e.g., [5]). These methods promise to shed considerable light on existing findings owing to their vastly improved spatial and temporal precision. However, they are not widely available to most language researchers.

fMRI

Neural activity in different brain areas can be indirectly measured via the local increase in blood oxygenation as participants understand or produce language. In the language network, but not in the multiple-demand network, blood flow appears to increase when the input is unpredictable [128]. The spatial resolution of fMRI allows inferences regarding the sources of neural activity and, when combined with functional localization

approaches [189], allows conclusions to be drawn regarding the underlying cognitive processes (e.g., language-specific vs cognitive control). The sluggish nature of the fMRI signal has prevented fMRI from being as widely used in investigations of linguistic prediction, relative to eye-tracking and EEG, but recent advances in analytical techniques have somewhat mitigated these issues.

Box 3.

Investigating prediction in language by measuring human and machine behavior

Eye-tracking in the visual world paradigm (VWP)

The coordinates of a listener's eye-gaze are recorded as they observe a visual scene or display while listening to a description of, or an instruction regarding, said scene. How often the participant's gaze lands on a depicted object is roughly proportional to how much the same or a related concept is active in the listener's mind. The listener's gaze often lands on images corresponding to the predicted continuation of the sentence (e.g., more looks to the cake after hearing 'the boy will eat the …' than 'the boy will move the …' [18,19]).

The VWP readily allows researchers to probe the nature of what is predicted by the listener in a given context, as opposed to indicating the magnitude of their response to a violation of prediction (as in ERPs). However, the set of available options for display requires careful consideration because it may bias what listeners look at, and the signal (consisting of fixations and saccades) is the discrete downstream consequence of myriad cognitive processes which are not yet fully understood [190].

Reading time (eye-tracking and self-paced reading)

As people read a sentence, they are likely to slow down on words they find more 'difficult', such as unpredictable words [16]. This slowdown can be measured by asking participants to read a sentence one word or phrase at a time – incrementally revealing the next word/phrase by button-press. This method is widely accessible to researchers and has the benefit of being directly related to the real-life impacts of linguistic prediction (e.g., how the features of a text affect struggling readers). However, this self-paced reading method has low temporal resolution. Using eye tracking to measure the time spent looking at a word (or words) affords greater precision, but both approaches are primarily one-dimensional – they index the duration of processing time – making it challenging to disentangle underlying mechanisms unless they are combined with other measures/approaches.

Sentence completions

Participants are asked to complete a sentence (e.g., 'In English class, the student read a …'). This is commonly referred to as a 'cloze task' [191]. When the next word is highly predictable, a majority of people will use that word (e.g., book) to complete the sentence (by definition). When the context does not allow a strong prediction ('At recess, the student read a …'), the distribution over possible completions will be closer to uniform (e.g., flyer, text, book, etc.).

Language models

Artificial neural networks and other language models (e.g., *n*-gram models which compute word probability from $n-1$ preceding words based on corpus frequencies), as well as distributional semantic models (vector-based representations of word meanings

based on their co-occurrences with other words or contexts), can be used to select stimuli for experiments and can be particularly useful when collecting norms from human participants is complicated or insufficiently precise [27]. However, probability measures from participants and language models do not always align, and how this alignment is related to model architecture remains poorly understood (Box 4 for further discussion).

Box 4.

What can we learn about prediction from neural language models and brain data?

A potentially promising direction for understanding prediction at the process level is the use of large artificial neural networks as mechanistic models rather than simply as statistical tools for computing surprisal values (or other metrics). Some researchers argue that, in the absence of relevant animal models, language models are the closest approximation to model systems for human language (discussed in [4]). In particular, approaches that compare specific architectures and probe the causal role of their components may help to adjudicate between classes of mechanistic proposals. Many such investigations find that models engaged in (hierarchical multiscale) prediction provide the best fit to brain data [4,60,62,192].

However, prediction may not uniquely account for the fit of language models to brain activity. Metrics such as the generality of representations (i.e., how well they transfer to a different task) [193] are also correlated with the ability to predict brain activity. Moreover, caution is warranted in interpreting these statistical relationships between neural language models and brain data. The ability of an artificial neural network to predict neural and behavioral data cannot be, on its own, taken as evidence that the model is mechanistically equivalent to human cognition (‘correlation does not imply cognition’ [194]; *cf* [195]). Deeper probing of model behavior and representations, their relationship to neural data ([196] for one approach), and the use of theory-driven benchmarks will be important for making stronger claims of equivalence.

In addition, smaller-scale models, operating exclusively over toy languages and that are constructed based on specific theoretical and neurobiological considerations, may also play an important role. What they lack in broad coverage and accuracy (e.g., of surprisal estimates) they make up for in interpretability. With small neural network models, researchers can exhaustively probe the internal representations and dynamics, thus providing complementary constraints for mechanistic proposals, and investigate questions that large language models – optimized for typical natural language processing tasks – do not address, such as the relationships between different brain networks (e.g., cognitive control and language).

Box 5.

Domain-general processes in language

A long-standing question in the cognitive science of language concerns the extent to which aspects of language processing involve domain-general processes. The exact meaning of 'domain-general' is often ambiguous. Plausibly, despite its functional specialization for language, the language network shares computational properties, such as prediction, learning from error, and inference, with most other regions of association cortex (e.g., [122,197]). However, it is unclear to what extent these computations occur within local circuits [182] or whether they are mediated by a hub [198] that is responsible for cognitive control across domains [199].

Independently, aspects of cognitive control have long been proposed to be involved in language processing [200]. These two threads are difficult to disentangle: some accounts view canonical computations and cognitive control as two sides of the same coin [135] or cognitive control as emerging from canonical computations at the highest level of a multi-level cortical hierarchy [201,202].

Cognitive control and associated terms such as executive function, working memory, and attention broadly refer to the ability to flexibly process information in a task-relevant way (e.g., [129,203,204]). The relevant computations appear to recruit a network of frontal, cingulate, and parietal brain areas (e.g., [130,205]) which are engaged regardless of the substrate of information processing (although some functional biases, such as toward visual vs auditory modalities, may exist within the network [206,207]). Although some correlational studies have suggested that cognitive control is involved in language processing based on the observation that individuals who perform well on cognitive control tasks are more likely to perform well on measures of language processing (e.g., [208,209]; *cf* [210]), these conclusions are often complicated by the challenge of measuring language abilities in a valid and reliable way [106,131]. On the other hand, during naturalistic language comprehension (i.e., without an explicit task), the frontoparietal cognitive control network appears to be minimally engaged [211,212]. Although 'passive' comprehension may not require cognitive control – and therefore the core functions of language processing may be independent of executive functions – many aspects of everyday language use are far from 'passive'. For instance, producing language requires the selection of a particular form to convey a meaning (among multiple options) that will be most appropriate for the communicative goal at hand [213], and real-world conversations often require compensation for 'noise' in the input and maintenance of multiple contexts on different time-scales (e.g., interpreting language according to the goals of different speakers, or combining local/sentence context with larger discourse context). Further exploration of language use in varied ecological contexts (e.g., in noise, with multiple speakers, etc.) and integration of models of cognitive control with language (e.g., [214]) may spur more specific mechanistic predictions about when and how language comprehension invokes cognitive control.

Outstanding questions

How do socio-contextual inferences impact predictions? For instance, a speaker's dialect can trigger a host of inferences about the language use and goals of a speaker, but how we flexibly adapt our predictions to the speaker context is unknown.

Do people predict more strongly when predictions are repeatedly confirmed than when they are disconfirmed? Indeed, people may engage more in prediction when individuals repeatedly produce predictable utterances rather than unpredictable utterances. However, several studies have suggested limits on whether people engage in 'rational adaptation' of their predictions.

Is the N400 ERP a reflection of prediction or updating? Neural network simulations show that N400 amplitude can be successfully modeled as a process of updating a probabilistic representation of meaning. However, experiments have thus far not supported a direct link between N400 amplitude and updating.

Is prediction sufficient for understanding? The purpose of human communication is plausibly to exchange thoughts between minds, potentially in the service of performing actions that are adaptive, and this requires understanding the meaning of what is being communicated. The relationship between prediction and understanding is not yet clear. For example, artificial neural networks can predict linguistic sequences with high accuracy, but it is unclear whether this endows them with the ability to understand.

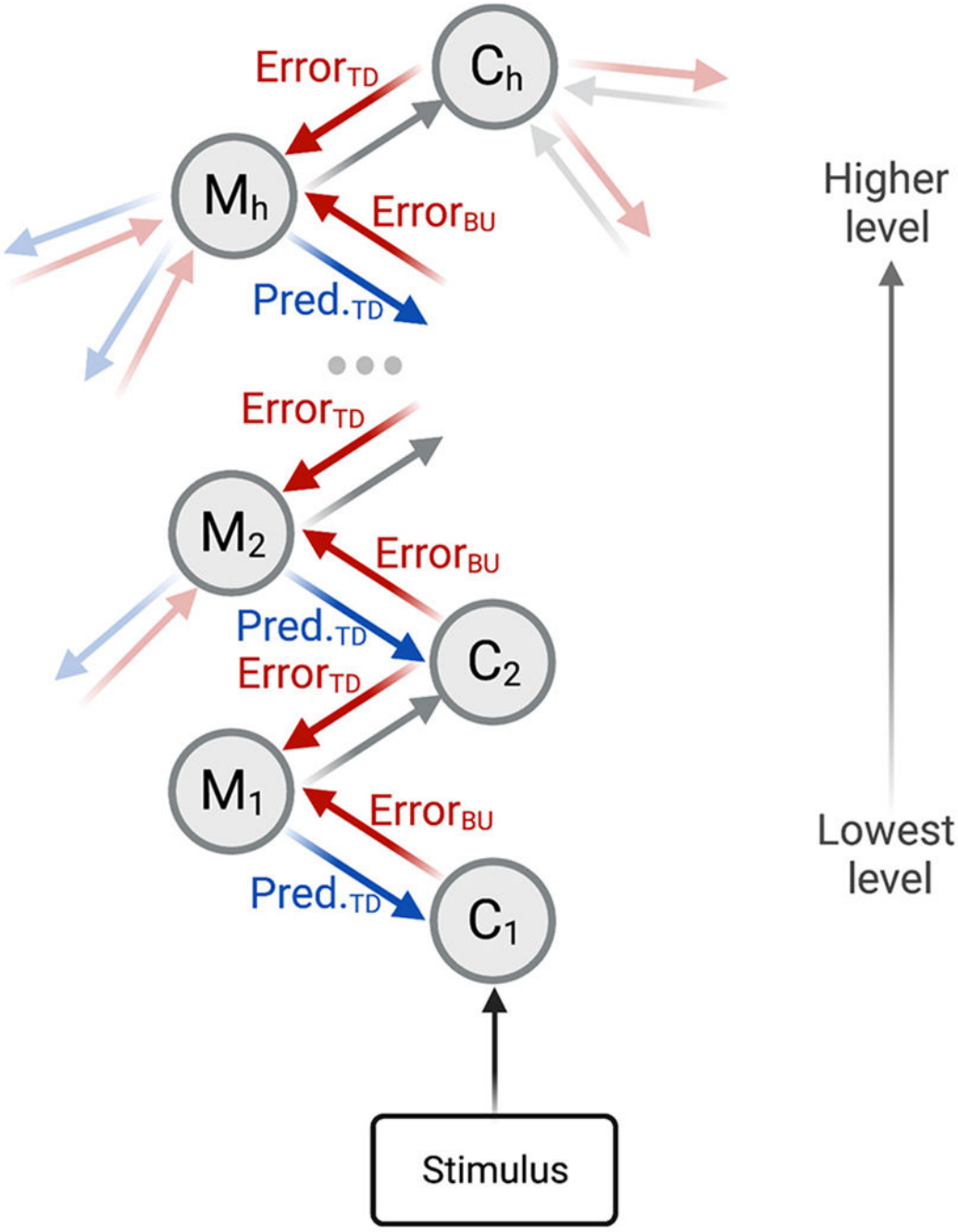

Figure 1. Schematic of predictive coding following Rao and Ballard [10].

M nodes indicate neural activity encoding a *model* of the level below. These models generate top-down (TD) predictions (Pred., blue arrows) about the level below. C nodes indicate comparisons between the activity predicted by the M of the level above and the actual activity at the current level. These comparisons result in top-down (TD) errors (red arrows pointing down) which adjust the estimate of activity at a given level and, crucially, bottom-up (BU) prediction errors (red errors pointing up) which serve to improve model

predictions at the next level up. Nodes at higher levels receive inputs from multiple instances of prediction errors at lower levels. Figure created with BioRender.com.

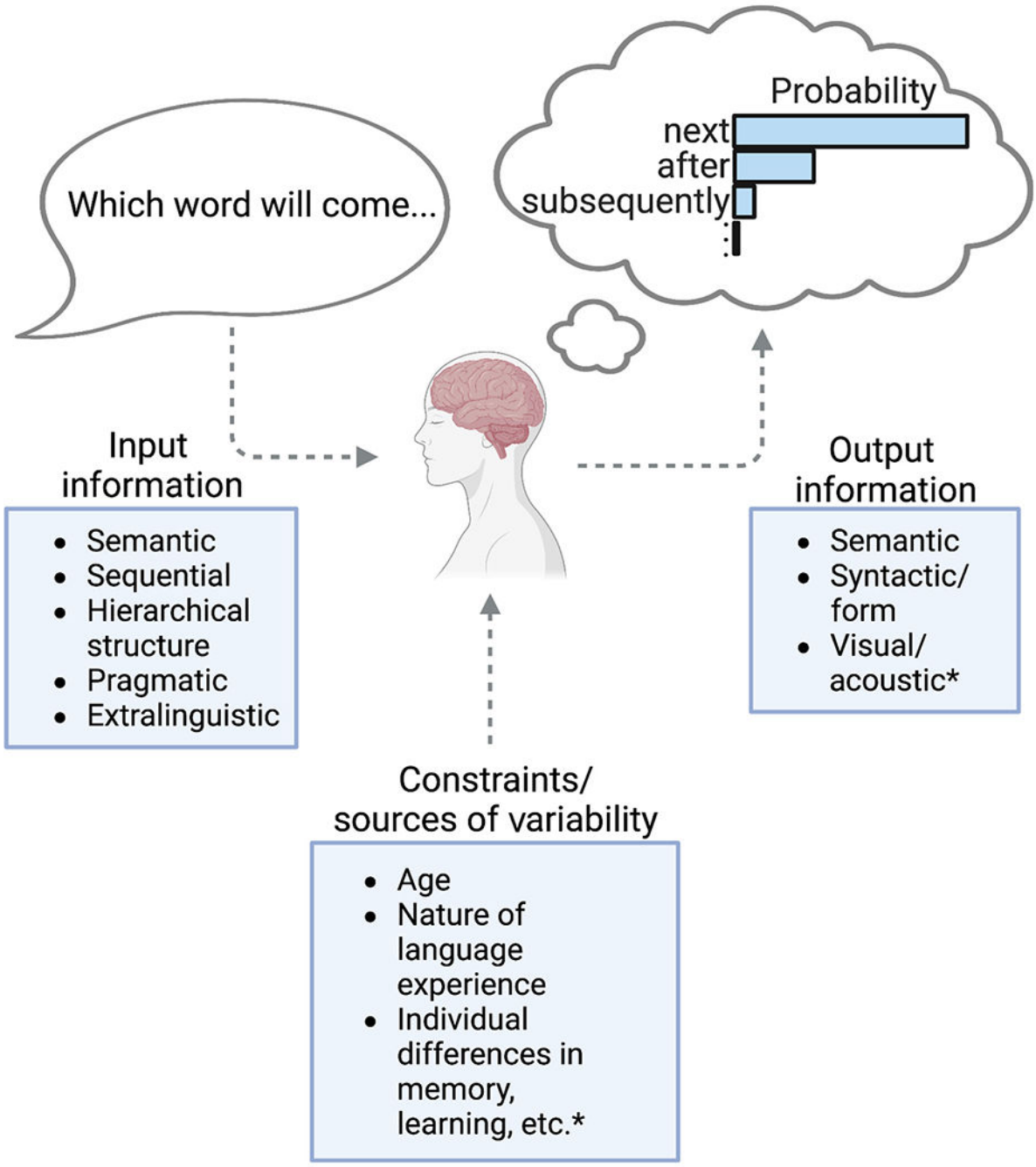

Figure 2. Prediction in language comprehension at the computational level.

Schematic summary of inputs to, outputs of, and constraints on the computation of prediction during language comprehension. An asterisk (*) indicates areas of ongoing debate/investigation. Figure created with BioRender.com.

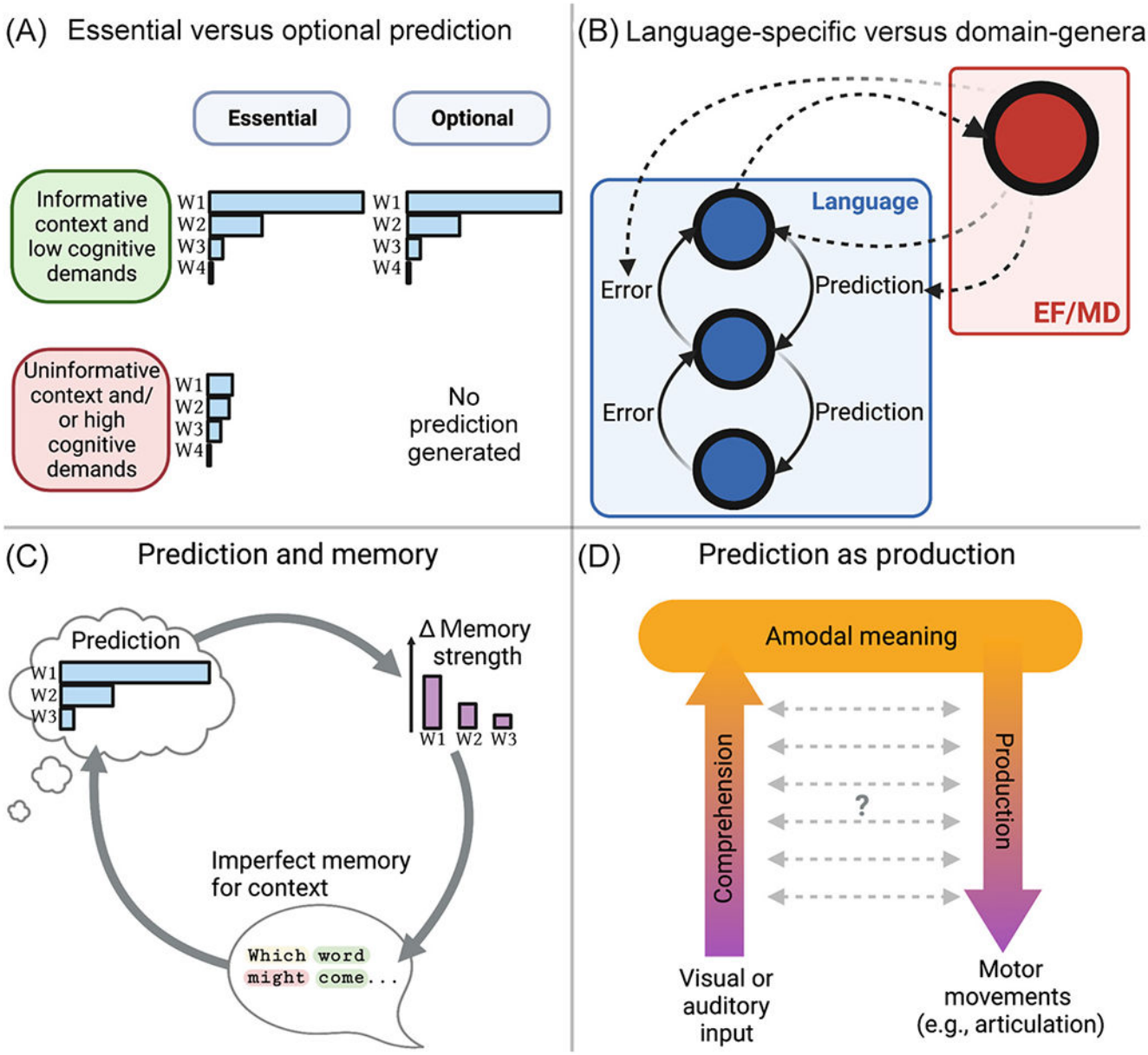

Figure 3. Unresolved questions that will inform mechanistic accounts of prediction during language comprehension.

(A) Essential versus optional. Blue bars schematically represent the probabilities of upcoming words (w1 = word 1, w2 = word 2, etc.). Only four words are shown, but the probabilities are distributed over all possible continuations and encompass meaning, form, etc. On an account of prediction as an essential component of language comprehension, predictions have higher entropy when the context is uninformative and/or their cognitive resources are taxed (e.g., unnaturally fast presentation, dual task). On an account of prediction as an optional strategy, no probabilities are computed under those circumstances. (B) Language-specific versus domain-general. Blue circles represent language network processes. Red circles represent multiple-demand network processes. Some aspects of linguistic prediction are likely implemented in local circuits (blue circles and arrows

reflect a predictive coding-like architecture in which predictions travel up and errors travel down the hierarchy). Domain-general cognitive control may play a role under specific circumstances (e.g., when there are errors in the input or when flexible switching between contexts is required), and the locus of its influence is unknown (as indicated by dashed arrows connected to either errors or predictions). (C) Prediction and memory. The three stages of the cycle represent (clockwise) (i) the generation of probabilistic predictions in response to language input (blue bars represent probabilities), (ii) the change in memory representations that results from the prediction that was generated and its relationship to the received input (e.g., a highly predicted input that is disconfirmed by the received input is still 'boosted' in memory), and (iii) the lossy memory for the context used to predict a subsequent input, which is determined by the state of an individual's memory for language. (D) Prediction as production. Comprehension and the prediction processes that support it receive sensory/perceptual inputs and extract meaning. Production consists of transforming the intended meaning of a speaker into motor actions. Although they tap into the same meaning representations, the low-level input/output representations of production (i.e., motor movements of speech articulation, writing, signing) and comprehension (vision, audition, touch) are distinct. Whether prediction during comprehension involves forward-simulating production, and how far down the production pipeline this simulation might go, is an open question. Abbreviations: EF, executive function; MD, multiple-demand system. Figure created with BioRender.com.