A deep hierarchy of predictions enables assignment of semantic roles in real-time speech comprehension

Yaqing Su[1,2], Itsaso Olasagasti[1,2*], Anne-Lise Giraud[1,2,3*]

1-Department of Fundamental Neuroscience, Faculty of Medicine, University of Geneva, Geneva, Switzerland
2-NCCR-EvolvingLanguage
3-Institut Pasteur, Université Paris Cité, Inserm, Institut de l'Audition, F-75012 Paris, France
* Joint senior authors

Abstract (216 words)

Understanding speech requires mapping fleeting and often ambiguous soundwaves to meaning. Humans are known to exploit their capacity to contextualize to facilitate this process, but how internal knowledge is used and deployed in real time remains an open question. Existing models of speech processing focus on either word recognition irrespective of meaning or interactions among abstract linguistic representations without time constraints, providing only partial insights into the dynamics of speech comprehension. Here, we present a model that incrementally extracts multiple levels of information from continuous speech signals in real time, based on the inversion of a generative model that represents the listener's internal knowledge of linguistic and non-linguistic processing levels in a nested temporal hierarchy. In each hierarchy, the model periodically incorporates bottom-up incoming evidence to update its internal representations and generate new top-down predictions. We show that a context level, beyond linguistic representations, can provide the model with semantic predictions informed by sensory inputs, crucial for the disambiguation among multiple meanings of the same word. We also show that hierarchical predictions can reduce peripheral processing effort via minimizing uncertainty and prediction error, especially when sensory precisions become degraded. With this proof-of-concept model we demonstrate that the deployment of hierarchical predictions is a possible strategy for the brain to utilize structured knowledge dynamically for speech comprehension.

Introduction

Understanding speech is a non-trivial feat. To extract information from ever-changing acoustic signals, our brains have to simultaneously "compress and recode linguistic input as rapidly as possible" for multiple representation levels (the "now-or-never" bottleneck (Christiansen and Chater, 2016)), while also keeping information in memory as we incrementally build up the meaning of an utterance (Tanenhaus et al., 1995). No computational framework to date has captured the transformation from continuous acoustic signal to abstract meaning: most speech processing models focus on either the lower-level recognition from acoustic to lexicon (Levinson, 1986, Mcclelland and Elman, 1986, Norris, 1994, LeCun and Bengio, 1995, Friston et al., 2021), or the higher-level linguistic manipulations without taking into account the constraint of elapsing time (Elman, 1990, Griffiths et al., 2007, Levy, 2008, Martin and Doumas, 2017, Friston et al., 2020).

In addition to the challenge of fleeting time, speech signals are often ambiguous and imperfect. However, humans exhibit extraordinary flexibility in making sense of ambiguous speech. We constantly make inferences based on our linguistic and nonlinguistic knowledge such as speaker identity and semantic context. The influence of internal knowledge on speech perception takes place

at all processing levels, e.g. filling the gap when the acoustic details are obscured in an utterance (Warren, 1970, Assmann and Summerfield, 2004, Sohoglu et al., 2012, Leonard et al., 2016), or interpreting a sentence containing semantically ambiguous words (Swinney, 1979, Rodd et al., 2005). Understanding how internal knowledge, both linguistic and nonlinguistic, is integrated with external input on the fly is key to deciphering speech processing in the brain, and explaining human's extraordinary flexibility in speech comprehension.

With the development of powerful neural networks (Devlin et al., 2018, Radford et al., 2019, Brown et al., 2020), it is now possible for a model to implicitly learn structured linguistic knowledge from an immense amount of written text, and apply such knowledge in language tasks such as coherent text generation. Despite their remarkable achievements in specific language tasks, these models are very resource-demanding and often make egregious errors showing that their performance is not rooted in human-like understanding of the language content (Floridi and Chiriatti, 2020). Being trained and evaluated on tasks involving predicting the next input, e.g. a word, it is virtually impossible for them to capture the abstract processing necessary for human language comprehension extending beyond linguistic forms (Bender and Koller, 2020). A key aspect of speech understanding consists in applying structured internal knowledge to extract relevant information from the input signal. How and what internal knowledge is deployed depends on the listener's behavioral goal, which can range from "understanding the message intended by the speaker" during a conversation, "gaining useful knowledge" from a lecture, to simply "predicting the next word" during an experimental task. A language model exploiting built-in linguistic as well as nonlinguistic knowledge, and driven by a behavioral goal, may hence be more powerful and polyvalent than one based on recognition and short-range prediction.

A recent study from Friston et al. (Friston et al., 2020) has applied these principles in a model of linguistic communication, in which a simulated "questioner" adaptively asks questions to an "answerer" to infer hidden information about two icons. Both conversers share the same two-level probabilistic generative model, where values (states) of several high-level "conceptual" components such as *semiotic* generate the low-level syntax and semantic states of each sentence, which in turn generate the corresponding word sequence. Components of this model, however, are independent within each hierarchy, therefore semantic and syntax information only influence each other at sentence onset/offset via higher-level components. As a result, the inverse (perception) model does not use each incoming word to incrementally update the belief of conceptual states or to change a semantic state inferred from a previous word.

Here, we establish a computational framework that applies abstract linguistic and contextual knowledge to incrementally extract multi-level information from the continuous speech signal. The model understands single sentences by assigning appropriate values to semantic roles and making reasonable judgements about the non-linguistic context of the utterance. Such process does not rely on an explicit representation of a full (Martin and Doumas, 2017, Martin, 2020) or partial (Friston et al., 2020) sentence, but on a probabilistic generative model that uses its linguistic and nonlinguistic knowledge to compose sentences in an incremental fashion. The generative model has a top context level that determines values of 2nd-level semantic roles, which are translated into a 3rd-level lemma sequences via linearized syntax rules. Each lemma produces a sequence of continuous, bottom-level spectro-temporal patterns via two intermediate hierarchies of finer timescales, integrating a biophysically plausible syllable recognition model (Hovsepyan et al., 2020). Unlike in Friston et al. (Friston et al., 2020), context and semantic states are maintained throughout the sentence but interact at the lemma rate, allowing the inverse model to modify previous estimates of these states with incoming evidence. During model inversion, each hierarchy sends prior estimates (predictions) at the rate of its subordinate level, and adjusts its posterior estimates using the discrepancy between the prior and the received input before sending down the next prediction. In this way, top-

down (internal knowledge) and bottom-up (input) messages are organized in a deep temporal hierarchy and alternate as the speech signal unfolds, providing a possible solution to the "now-or-never" bottleneck (Christiansen and Chater, 2016) that is also consistent with the predictive coding hypothesis of perception (Rao and Ballard, 1999, Friston, 2009, Clark, 2013).

With a small scope of knowledge adapted from stimuli in MacGregor et al. (2020), the model can extract contexts and semantic roles from ongoing speech signals and resolve semantic ambiguity using new information. We show that informative top-down prediction reduces processing time and energy cost in speech perception compared to uninformative predictions, but overconfidence in prediction also compromises inference accuracy. In addition, the linguistically informed model structure allows for hierarchy-specific metrics for the characterization of inter- and within-hierarchy computations.

This proof-of-concept model demonstrates a possible computational scheme of speech processing in the brain in which top-down prediction serves as a key computational mechanism for information exchange between hierarchies, driven by the goal of comprehension. Although we cannot draw any direct correspondence between the model's computational motifs and its possible neurophysiological implementation, correlations between model-derived metrics and neural responses may provide insights into the functional roles of various neuronal signals during speech perception.

Results

A deep hierarchical model of speech comprehension

We developed a model of speech processing based on the idea that the goal of the listener is to understand the message conveyed by an utterance. Appropriate understanding entails retrieving useful information from the utterance and optimally mapping it to the listener's knowledge of the world, not restricted to linguistic representations (Figure 1A). Our model of the listener's internal knowledge therefore consists of two parts that are both implemented in the form of probabilistic generative models. The first part exemplifies knowledge about the world by defining events and properties that are constrained by specific contexts. This knowledge reflects both the listener's own beliefs about the world that do not depend on specific speakers, and its beliefs about how the current speaker composes a sentence to express a message, but we do not make an explicit distinction between the two here. For example, under the context of a tennis game, the listener knows (that the speaker knows) about special winning serves, about runs to return a ball etc. The serve or the run may be the central role in an event of winning a game, or described as having a certain property (e.g.: being surprising). Under a different context of a poker game, the listener would know some special cards in the deck that can also be part of an event or entail some property. The 2nd part of the model converts these events or properties into linguistic forms by choosing between a number of possible lemmas in an appropriate order, e.g. the special winning serve can be expressed as a single word "ace" early in the sentence, and finally into spoken utterances in the form of spectro-temporal sound patterns via a deep temporal hierarchy (Figure 1B). These two parts are hierarchically linked via semantics and syntax. The inversion of this generative "world knowledge" model fulfills the mapping from the sound patterns to abstract semantic roles and contexts by estimating the probability of every possible value (state) of each element (*factor*) in the knowledge hierarchy (Figure 1A), thus providing the listener with the means to understand the utterance produced by the speaker.

In all, the model includes five levels, each consisting of several factors represented in rectangles in Figure 1A, and with multiple possible values (states) listed in Table 2 in Methods except for the *acoustic* factor. Probabilistic mappings and transition probabilities between these values are defined in Methods and Appendix. The final output of the generative model (i.e. the input to the perception model) is the continuous spectro-temporal pattern of the speech signal sampled at 1000 Hz and divided into six frequency channels (see Methods). Lengths of stimuli are fixed: each sentence consists of 4 lemmas, each lemma of 3 syllables, and each syllable of 8 spectral vectors. Every spectral vector is deployed into 25ms of time-varying continuous signal, thus each syllable effectively has a duration of 200ms (Greenberg et al., 2003).

Next, we show how this model understands simple sentences and deals with semantic ambiguity, and demonstrate the role of top-down predictions in these processes. We assessed its performance with different sentence stimuli and parameter settings focusing on two aspects: 1) probability distributions that describe the model's beliefs (or predictions) about possible states over time, and 2) divergence and entropy, which summarize informational changes underlying the evolution of beliefs (see Methods).

Stimuli are adapted from (MacGregor et al., 2020) and illustrate the use of internal knowledge to disambiguate speech. All sentence stimuli in the following sections share the same structure (see Methods for a complete list of possible sentences):

One more [MIDDLE WORD] wins [END WORD].

The middle word can have either one unambiguous or multiple possible meanings, each meaning pointing to one context of the sentence. The end word either resolves the semantic ambiguity of the middle word or not. A disambiguating end word can also follow an unambiguous middle word without affecting its interpretation.

The use of knowledge about the world to interpret speech

We first test how the model processes speech stimuli, with a focus on the timing of the incremental estimation process at the top levels, where "meaning" is extracted by assigning values to semantic roles.

Consider the following two sentences, A: "one more ace wins the tennis" and B: "one more ace wins the game". Both sentences contain the ambiguous word "ace", which can be associated with a special serve in tennis or a special card in a poker game. The final word in the first sentence disambiguates "ace" to mean a special serve because "the tennis" can only be generated from a tennis game context, which applies to the whole sentence including the preceding "ace". In the second sentence, however, the ambiguity remains unresolved; the game can still refer to a tennis or a poker game. In the latter case, the interpretation of the word "ace" will depend on the listener's preference. In the simulations in this section, we introduce a prior preference for poker as a context to reflect the preference of the general population (MacGregor et al., 2020).

The word "ace" introduces ambiguity because it points to two possible states for *agent* ("tennis serve" or "card A"), each of which points to a separate state for *context* ("tennis game" or "poker game", Table 1 in Methods). Figures 2A and 2B show the evolution of the model's beliefs about context and semantic factors for the two sentences. The ambiguity is reflected in the posterior estimates of *agent* and *context* between the offset of "ace" and the sentence ending word, where the model assigned nonzero probabilities to "card A" and "serve" as the *agent*, and "poker game" and "tennis game" as the *context*, and near-zero probabilities for other states (Figure 2A). Probabilities for poker-relevant states were higher (darker colors) due to the contextual preference. The verb "wins" did not change the model's estimation for the *agent* or the *context*, but clarified the sentence *type* to be "event" and the *patient* to be nonempty, again with a preference towards poker. After the model heard "the tennis" (Figure 2A), it immediately resolved its beliefs of the *agent*, the *patient* and the *context* to the opposite of its prior preference. When the sentence ended with "the game", (Figure 2B) which didn't point to a clear resolution, the model followed its preference with enhanced beliefs as a result of the entropy reduction entailed by belief updating, but not as clearly resolved as with "the tennis" (see next section).

The results in Figure 2 demonstrate how prior knowledge and preferences can shape and guide the extraction of semantic roles and contexts from the speech signal as it unfolds. Such a process is influenced by prior knowledge not only in the perception of semantically ambiguous words, but also in the details of message passing that give rise to its estimates. Specifically, the amount of information maintained between belief updates as quantified by entropy, and the magnitude information change induced by an update as quantified by the Kullback-Leibler divergence, both vary with the model's prior preference for *context* states (Figure 3). Thus, individual preferences could influence the magnitude of belief updating (divergence) and information maintenance (entropy), quantities of interest since they are often correlated with neurophysiological signals in the interpretation of experimental data (see Discussion).

Perceptual ambiguity and disambiguation rely on different computational processes

As it is not straightforward to quantify individual subjects' preferences in their internal knowledge of speech processing, one common practice in neurophysiological experiment design is to contrast conditions with high versus low values of perceptually and computationally relevant variables, such

as word entropy and surprisal, that are conventionally derived from population language tests (DeLong et al., 2005, Wang et al., 2018, Mamashli et al., 2019, Rodd et al., 2005, MacGregor et al., 2020) and recently from computational models (Koskinen et al., 2020, Donhauser and Baillet, 2020, Caucheteux and King, 2022, Goldstein et al., 2021, Heilbron et al., 2021, Schrimpf et al., 2021). However, interpretations of neurophysiological correlates to these computational variables are often limited to phenomenology, in that a significant difference in the averaged neural response to the high vs. low condition is often directly attributed to the qualitative change in the corresponding variable without establishing what computational processes may give rise to the neural responses. Figure 4A and 4B were set up to match the example comparisons in the MEG study of MacGregor et al. (2020), where the authors identified an increase in the sensor-space response magnitude following the offset of a semantically ambiguous word compared to an unambiguous word ("ace" vs. "sprint"), as well as an increase following the offset of a word that resolved a previous semantic ambiguity ("the tennis" after "ace") versus either a different word that did not resolve the ambiguity ("the game" after "ace") or the same word but not resolving any ambiguity ("the tennis" after "sprint").

Figure 4A contrasts the inference process between sentence A [ACE-tennis] and sentence B [ACE-game] in Figure 2 using their derived information metrics ([ACE-tennis] minus [ACE-game]), focusing on the context, the agent, and the patient factors that were most affected by the set conditions. The entropy in the two sentences only shows a negative difference (lower entropy for ACE-tennis) at the context level at the sentence offset, indicating greater residual ambiguity in the context, but not in the agent or the patient, after hearing "the game". The divergence, however, shows a positive difference at the sentence offset, reflecting the reversal of prior preference upon hearing "the tennis". Figure 4B displays the differences in entropy and divergence between sentence A [ACE-tennis] and C [SPRINT-tennis]: "one more sprint wins the tennis" ([ACE-tennis] minus [SPRINT-tennis]). In the model's knowledge, the word "sprint" has an unambiguous meaning of "fast run" and unambiguously points to the context of a tennis game. Compared to "sprint", "ace" introduced higher entropies for all three factors, reflecting the model's uncertainty about their states. The uncertain beliefs after hearing "ace" were also less different from those of the previous time point compared to the unambiguous "sprint", resulting in a negative difference in divergence. After hearing the same sentence-ending lemma "the tennis", the model was highly certain about states in both sentences, and the difference in entropies became minimal. Sentence A entailed higher divergence because here "the tennis" was different from what the model expected from its preference.

Although it seems intuitive that ambiguity would result in an increase in entropy, and disambiguation (towards a less preferred candidate) in an increase in divergence, one cannot conclude that neurophysiological signatures in response to ambiguity or disambiguation, especially those in sensor space, respectively reflect the computation of entropy or divergence. As can be seen in Figure 4, a difference in entropy between two conditions is often associated with a difference in divergence but in the opposite direction, with magnitudes varying across hierarchies and across factors within the same hierarchy. Thus, a direct link from computational processes characterized by entropy and divergence to aggregated sensor space response such as ERP, and further to cognitive phenomena of semantic ambiguity and disambiguation, cannot be established because both phenomena involve a complex combination of computational processes of different types and hierarchies. Such complexity is in line with the finding of MacGregor et al. (2020) that the two sensor-space phenomena were localized to different but overlapping sources. With an appropriate upscaling of the model to capture the internal knowledge of real-life subjects, it is plausible to further dissociate different computation processes by correlating model-derived information

metrics, importantly at different hierarchical levels and factors, with source-, time- and frequency-specific responses (see Discussion).

Top-down prediction reduces processing effort

The model works by iteratively calculating the discrepancy between top-down predictions (expectation of the input) and bottom-up input at each hierarchy and using such discrepancy to modify the state estimates of superordinate hierarchies. This does not imply that the model needs to make the best prediction for the next input: hierarchical predictions are a necessary computational mechanism in the relaying of information, which contributes to making correct inferences even if the actual input does not correspond to the predicted one. To examine how the content of prediction may influence the inference process, we compared the model responses to "one more ace wins the tennis" with informative (best estimation of the next input) versus uninformative (uniform distribution across all possibilities) top-down predictions. We found that the prediction scheme can influence both the time course and the final estimate of the model.

Figure 5A and 5B respectively show top-down priors and posterior estimates for lemma and syllable levels with the same parameters as Figure 2A, i.e. with informative top-down prediction. The predictions reflect both prior knowledge and the updated estimation of superordinate levels, and posterior estimations immediately converged to the correct states after receiving the determining input that unambiguously points to a specific state, for example the second syllable in the last lemma. When top-down predictions were made uninformative (Figure 5C), the model still made correct inferences about every input, but with a slight delay for syllables (note the short vertical bars at the beginning of some syllables in the bottom panel of Figure 5C, indicating slower convergence to the estimate compared with Figure 5B). Figure 5D and 5E quantify the difference between the two conditions by contrasting the entropy and cumulative divergence during the inference process. Unsurprisingly, informative predictions lead to reduced processing effort as quantified in terms of entropy (maintenance of possible items) and divergence (magnitude of updates after the integration of new evidence). Despite the differences at lower levels, the model response at the semantic and context levels are nearly identical in these two conditions because the model reached the same, almost-certain lemma estimates at the time of semantic updating (at each lemma offset), and the updating was instantaneous in both cases due to the small number of total possibilities (not shown).

So far, we have simulated the model with the ideal scenario of arbitrarily high precisions (see Methods) at the bottom continuous level. The precision parameter is a weight applied to the discrepancy between predicted and received input that contributes to the update of the posterior estimate (Friston et al., 2008). In general, a high precision implies that fine details from the input are utilized to evaluate the mapping between the input signal and the generative model, analogous to a perfect periphery that preserves the best possible spectro-temporal information from the acoustic input. It has been suggested that top-down predictions may be especially important under challenging situations, e.g. impaired auditory periphery or deteriorated speech signal (Sohoglu et al., 2012, Peelle, 2018). Figure 6 shows the comparison of informative vs. uninformative predictions similar to Figure 5, in a model with lower peripheral precisions. In particular, we lowered both the precision for the continuous states as well as for comparing the input with predicted activity in the six frequency channels (see Methods), analogous to lesioning the local computations supported by lateral connections and the cross-level information carried by bottom-up connections, respectively (Friston et al., 2008, Yildiz et al., 2013). Syllable identification was delayed in both cases when compared to their intact-periphery counterparts (Figure 6A vs. 5B, 6B vs. 5C), and the delay was more pronounced with uninformative predictions. This dramatic delay is accompanied by higher entropy (red curve in Figure 6C), which showed both a higher maximum and a longer decay for every

syllable. Having to make more effort in recognizing each syllable, however, may not be completely undesirable: in Figure 6A, the model relied on its prior knowledge, saving processing effort but ended up recognizing the last lemma incorrectly as "the poker". The tradeoff between processing effort and accuracy has been well-documented in the decision-making literature (Payne et al., 1988) and later recognized by neuroscientists as neuroeconomics (see review by Eckert et al. (2016) on listening), which reveals that humans flexibly adapt their strategy in challenging scenarios where high accuracy and low effort cannot be achieved together. Our results suggest that such tradeoff can be achieved via adjusting the content of top-down predictions and one's reliance on them, an ability likely lacking in certain neuropsychological disorders (Parr et al., 2018). The comparison between Figure 5 and 6 also indicates that listeners with unimpaired periphery may choose their prediction strategies to optimize one aspect without compromising the other as much as the impaired.

Overall, the model demonstrates that hierarchical prediction, whether highly informative about the next input or not, can serve as a key computational mechanism for robustly extracting structured information from ongoing speech, and that informative predictions are desirable when processing effort needs to be minimized. With an impaired periphery, greater effort is required to obtain an accurate perception.

Discussion

The idea that our brains adaptively entertain internal models to facilitate language comprehension underlies most current research in speech (language) perception. Nevertheless, how internal knowledge is deployed in time is an open question, and may be key in resolving the form-meaning discrepancy in neural-network language models (Bender and Koller, 2020). Here, we attempt to establish a foundational framework that dynamically exploits general knowledge in speech comprehension to bridge this gap. We implement the listener's internal knowledge as a probabilistic generative model that consists of a non-linguistic general knowledge (cognitive) model and multiple temporally organized hierarchies encoding linguistic and acoustic knowledge. Speech perception, modeled as the inversion of this generative model, involves interleaved top-down and bottom-up message passing in solving the computational challenge of extracting meaning from ongoing, continuous speech. We show that the model makes plausible inference of hierarchical information from semantically ambiguous speech stimuli and demonstrate the influence of prior knowledge on the inference process. We also show that hierarchical predictions can be exploited to reduce processing effort. The model tries to mimic human language comprehension by implementing incrementality and prediction (Altmann and Mirkovic, 2009). It can be potentially expanded towards a comprehensive model of natural language **understanding**, and guide the interpretation of neurophysiological phenomena in realistic listening scenarios.

Language comprehension as semantic role assignment

Although we emphasize that speech (language) comprehension is driven by high-level behavioral goals, to achieve comprehension the appropriate assignment of semantic roles conveyed in the utterance is a necessary step. Semantic roles can be viewed as an interface between linguistic and nonlinguistic representations, the latter being a more abstract and fundamental format of our internal rendering of the world that can be expressed in multiple modalities not restricted to verbal language. The process of semantic role assignment is fundamental in psycholinguistic process theories (Tanenhaus et al., 1989, McRae et al., 1997, Altmann, 1999), yet seldom reflected explicitly in existing computational models of language (cf. (Altmann and Mirkovic, 2009)). A major challenge for modeling semantic role assignment during language processing is in combining meaning extraction with compositionality: words that carry semantic contents are presented in an order dictated by compositional rules, thus the extraction of persisting meanings must take place dynamically alongside the decomposition. These two aspects have only been addressed separately in some existing models, e.g. topic models (Blei et al., 2003, Griffiths et al., 2007) fulfill (lexical) semantic processing as they extract the "gist" in a collection of words but ignores the word order, whereas the Discovery of Relation by Analogy model (Martin and Doumas, 2017, Martin, 2020) learns the time-based binding rules that decompose words and phrases into hierarchical structures but does not have any explicit representation of semantic knowledge.

A recent model of linguistic communication (Friston et al., 2020) did incorporate abstract non-linguistic (geometric) knowledge and compositionality, but lacked the incremental nature of the meaning-building process in humans (Tanenhaus et al., 1995). The generative model had access to a set of possible formats of complete sentences and a set of properties that described simple geometric objects. By applying nonlinguistic knowledge under the goal of resolving the object properties, the model generated sentences by picking the most probable sentence format and filling specific positions with the most helpful descriptive words. The inverse model thus comprehends a word sequence by inferring the sentence format and capturing keywords at the corresponding positions. This template-matching strategy realized a form of meaning-structure conjunction. However, it constrains the model comprehender to update its estimate of the sentence at the sentence offset instead of on the fly during the sentence.

Our model achieves human-like speech (language) comprehension in that it applies syntactic rules to dynamically update values assigned to semantic roles with each incoming lemma. It does not have a direct representation of sentences, but incrementally builds up its understanding of an utterance through incorporating new evidence (here the speech signal) into current beliefs of semantic roles. This feature is enabled by the construction of a hierarchical generative model that explicitly represents meaning and links semantic roles with syntactic elements. Although the model manages very limited linguistic and world knowledge, i.e. a small number of possible states for each factor, it can in principle process any plausible sentences within this miniature world given that the corresponding probabilistic mapping matrices are set to reflect all possibilities. These probabilistic mappings in turn influence the model's perception through the information exchanged between processing hierarchies as is shown in Figure 2-4, which we believe plays an important role for linking the model's computational principles to neurophysiological data of speech information processing in the human brain.

Understanding neural information transfer through divergence and entropy

Brains process internal and external information with unparalleled efficiency. Two types of information theoretic metrics have been of particular interest in establishing the connection between abstract information and biophysical signals to probe the brain's information processing capacity: surprisal (related to, but distinct from divergence) and entropy.

The correlation between language processing (mainly reading) load with how surprising an incoming word is given its sentential context has been well-documented using behavioral metrics (Hale, 2001, Demberg and Keller, 2008). Efforts in associating neurophysiological responses to surprisal for next-word expectation, either based on cloze probability tests (DeLong et al., 2005, Wang et al., 2018, Mamashli et al., 2019, MacGregor et al., 2020) or the probabilistic distribution of predicted next-word input estimated by computational models (Koskinen et al., 2020, Donhauser and Baillet, 2020, Goldstein et al., 2021, Heilbron et al., 2021, Schrimpf et al., 2021, Caucheteux and King, 2022), largely credit Levy's influential work on expectation-based comprehension. Levy proposed a formal relationship between incremental comprehension effort and the Kullback-Leibler divergence (KLD) of syntactic structure inference before and after receiving a word input W, and proved that the KLD reduced to surprisal of W given the previous word string if the extra-sentential context remains unchanged. Although these studies robustly found neurophysiological correlates of word surprisal, focusing on this aggregated measure without explicitly modeling probabilistic representations above the word level may not be enough to tease out the influence of high-level factors on language processing (Figure 3) (cf. (Caucheteux and King, 2022)). High-level processes presumably explain conflicting findings across studies on evoked response (Kuperberg, 2007) and underlying neuronal circuits (MacGregor et al., 2020, Mamashli et al., 2019, Donhauser and Baillet, 2020) of word surprisal, because different experimental paradigms likely tap into different language processing modes, making word surprisal too coarse a measure. Here, we demonstrate the possibility to explicitly model information transduction above lexical processing and use KLD as a universal metric to quantify information transfer, in line with some predictive coding hypothesis that proposed KLD to be driving the prediction error signal transmitted between cortical hierarchies (Friston and Kiebel, 2009, Bastos et al., 2012). To adapt the model to different tasks, the factorization of the top level in our model, and its relationship with the semantic and context level, can be designed to reflect corresponding perceptual processes.

As for entropy, the measure of information in a system (Shannon, 1948) that represents the uncertainty or ambiguity in linguistic stimuli, has attracted less focus compared to surprisal metrics (Willems et al., 2016, Donhauser and Baillet, 2020, MacGregor et al., 2020, Gwilliams et al., 2020). This is partly because these two metrics are often correlated (e.g. Figure 4), and surprise as a marker

of instantaneous change can be more straightforwardly interpreted from neurophysiological paradigms that essentially work by detecting between-condition changes. However, there is little consensus on how information is maintained between two instantaneous belief updates, and entropy may be valuable in investigating such information *representation* in the brain. Intuitively, higher entropy implies greater effort to maintain more possibilities, and less precise estimates thus weaker influence in top-down predictions, but it is unclear what neural activities can reflect such effects. Noninvasive whole-brain imaging may inform us when and where the effort takes place given that entropy and divergence are properly dissociated using computational models (Donhauser and Baillet, 2020), but the biophysical implementation, e.g. firing patterns of neuronal ensembles, may only be revealed by invasive methods.

Let's point out that both KLD and entropy are measures of probability distributions within representational levels, which cannot directly stand for the information transmitted between levels. For inter-level information, it is possible to decompose the KLD into bottom-up prediction errors and top-down priors (Friston et al., 2017), and apply known neurophysiological probes, e.g. neural oscillations, to distinguish these two flows in the brain (Bastos et al., 2012, Giraud and Arnal, 2018, Giraud and Poeppel, 2012, Arnal and Giraud, 2012, Bastos et al., 2020, Hovsepyan et al., 2022) .

Insights for biophysical implementation by integrating neural oscillations

Although no definitive conclusion has been drawn on the anatomical circuits involved in message passing during speech perception, a converging view is that the extraction of different representational hierarchies is likely distributed in networks that perform multiple subprocesses in parallel (Egorova et al., 2013, Fedorenko et al., 2016, Pulvermuller, 2018, Fairs et al., 2021). Recent temporally and spatially resolved neuroimaging studies suggest that neural oscillations may be a candidate information transmission mechanism in these subprocesses. The discrete portion of our model, or in theory any model with explicit structural and timing information (Martin and Doumas, 2017, Martin, 2020), can provide a template for organizing distributed oscillatory activities into functional hierarchies through correlating latency- and frequency-specific neuronal dynamics with model-derived information metrics.

Specifically, oscillatory activities in the low-gamma (25-40 Hz) band package samples from fast variations in the acoustic input, which is then parsed into syllabic and phrasal (prosodic) information likely encoded as phase alignments of slower oscillations in theta (4-8 Hz) and delta (2-4 Hz) bands (Giraud and Poeppel, 2012, Arnal and Giraud, 2012, Ding et al., 2016, Lakatos et al., 2019, Hovsepyan et al., 2020, Rimmele et al., 2021). Whereas these bands appear to be entrained to bottom-up stimulus characteristics, they are also causally influenced by top-down modulations (Park et al., 2015) that may be coordinated by endogenous alpha (8-12 Hz) and beta (14-20 Hz) oscillations (Arnal and Giraud, 2012, Fontolan et al., 2014, Pefkou et al., 2017, Bastos et al., 2020) (also see (Murphy, 2018, Meyer et al., 2020) for integrative hypotheses). Moreover, the amplitude of slow oscillations as well as their interactions are shown to be modulated by surprisal and entropy in the speech signal at their corresponding linguistic timescales (Mamashli et al., 2019, Donhauser and Baillet, 2020, MacGregor et al., 2020). Thus, it is reasonable to associate time-locked evoked responses with model-identified informational changes in the brain's distributed representations, and further distinguish directional communication between neuronal sources of such evoked response by frequency channels. Whether this approach can be applied to low-level sensory processing requires further understanding in the neuronal interface between continuous and discrete (symbolic) operations, which are likely implemented in different fashion both anatomically and computationally (Felleman and Van Essen, 1991, Barrett and Simmons, 2015, Friston et al., 2017).

Limitations of the model and further development towards natural language understanding

In this work, we provide a basic model that integrates linguistic and world knowledge in speech perception, with a focus on resolving ambiguity in semantic role assignment within a reduced language and world model. For the application in modeling realistic speech comprehension and its neurophysiological correlates, two obvious limitations need to be overcome.

First is the model scale. Although we constructed the model's internal knowledge to support multiple possibilities in context/semantic assignment and (linear) syntactic composition during comprehension, this is far from the wealth of vocabulary and linguistic and nonlinguistic knowledge mastered by a real listener. A model with an adequately large parameter size may not qualitatively differ from a reduced model, but is necessary for quantitative analysis in relation to biological responses. A plausible algorithm of statistical parameter learning of structured contextual and semantic knowledge is the one proposed for the "topic" model of semantic representation (Blei et al., 2003, Griffiths et al., 2007), which finds topics and estimating prior distributions for the generative model of gist→topic→word from text corpora. Griffith et al. (Griffiths et al., 2007) also pointed to a possible way to integrate complex syntax and semantic generative models by replacing one component in a syntax model (Griffiths et al., 2004) with such a topic model. This would allow the syntax model to determine an appropriate semantic component for the current timepoint and the semantic model to generate a corresponding word, which is consistent with the way semantic and syntax factors interact in our current model. More recently, Beck and colleagues (Beck et al., 2012) showed that a formal equivalence of the topic model can be implemented via a probabilistic (neural) population code, providing another plausible path to upscaling the model.

Another limitation is that the model has a fixed length for each hierarchy. This issue has been partly addressed in Hovsepyan et al. (Hovsepyan et al., 2020), where a theta oscillator can adapt to the speech envelope, enabling the processing of syllables of different durations. To enable variable numbers of syllables per lemma, it may be necessary to separately model the previous and the current lemma, and let the first syllable of the new lemma signal the offset of the previous lemma as well as the onset of the new one. This is incompatible with the model implementation we adapted here (Friston et al., 2017), but realizable if formal relationships among factors are properly defined, such as in a recent Bayesian model that exploited lexical knowledge in syllable segmentation (Nabe et al., 2021). The problem of a fixed number of lemmas per sentence is automatically resolved with a complete syntax model. However, establishing formal relationships in linguistic processing or modeling nonlinear syntax within a Bayesian framework are both extremely challenging tasks.

Overall, this model adopts a different and complementary perspective from the rapidly developing world of large-scale natural language models (Devlin et al., 2018, Radford et al., 2019, Brown et al., 2020) in that it puts upfront the gross biological factors that motivate language in the first place (Hauser et al., 2002, Corballis, 2009, Greenfield, 1991, Fitch, 2012), rather than those that allows for matching human performance via selected measurements. This approach acknowledges that human language has emerged and evolved under evolutionary pressure that is both enabled and constrained by our specific biological substrates. We implemented these principles by explicitly including nonlinguistic components in the model architecture and using hierarchical (as opposed to aggregated) prediction as a general computational strategy. Although here we focus on a passive listener, a comprehensive model of human language understanding should also consider the action of language, i.e. language production and multi-person communication (Friston et al., 2020) where language serves as a medium to achieve shared goals (Galantucci et al., 2006, Hickok and Poeppel, 2007, Pulvermuller and Fadiga, 2010, Bender and Koller, 2020, Castellucci et al., 2022).

Summary

We present a computation framework for speech comprehension via extracting semantic roles from ongoing natural speech. By incorporating internal knowledge via hierarchical message passing, the model makes human-like incremental inference of semantic and contextual information conveyed in continuous speech signals. It demonstrates the perceptual effect of different degrees of top-down predictions and has advantages in accounting for neuronal activities underlying information processing beyond the word level. We provide strategies to apply the model in linking neural message passing with oscillatory mechanisms and point to possible directions for further development.

Figures

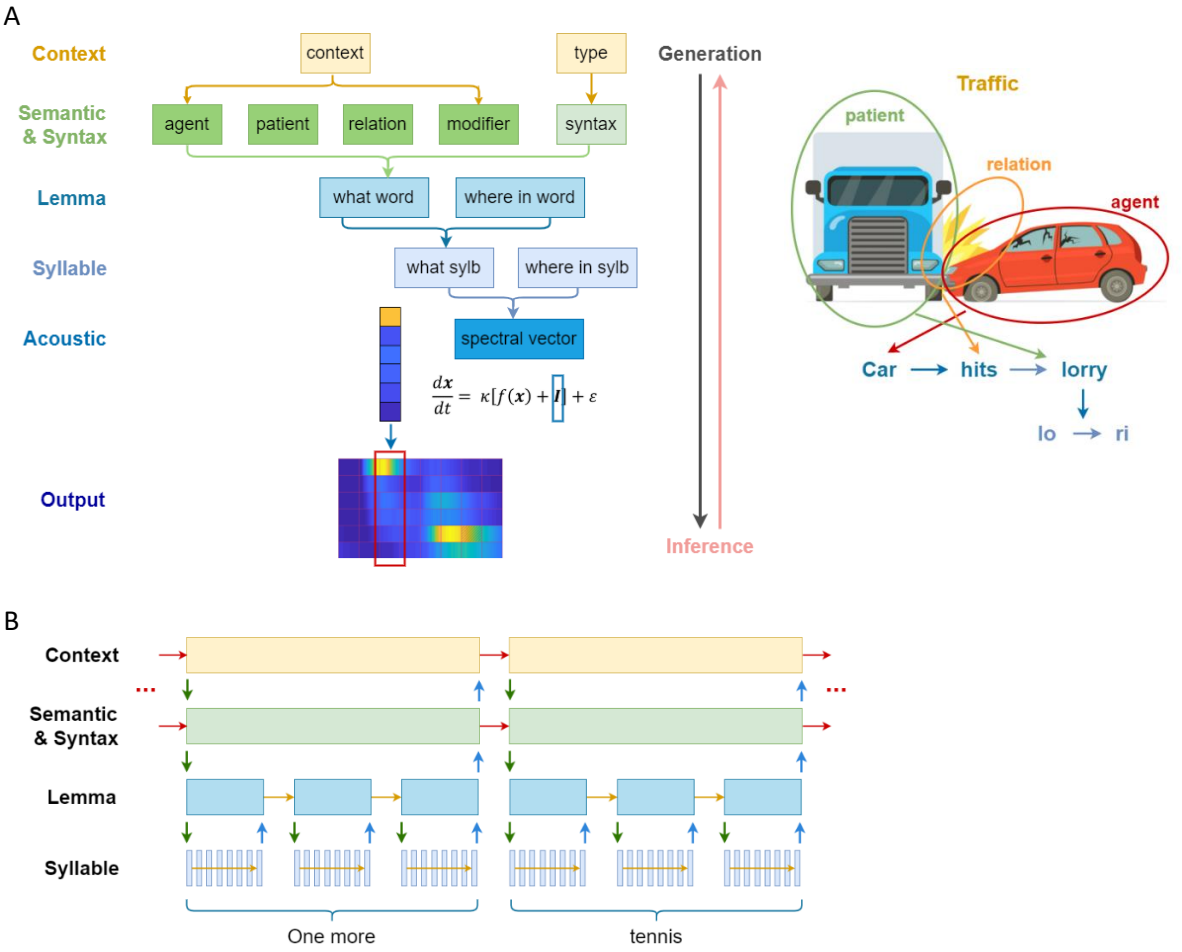

Figure 1. A generative model of speech and its inversion. 1A. Schematic of the generative model. Left: information conveyed in a speech signal is roughly separated into six hierarchies. To generate speech, the model first assigns values to semantic roles according to the contextual knowledge and determines a (linear) syntactic structure from the type of the message it's expressing. Together, semantics and syntax generate an ordered sequence of lemma units. Each lemma unit generates a sequence of syllables, which in turn generates a sequence of spectral vectors. Each spectral vector unit is then deployed as a continuous acoustic signal of 25 ms. Inference corresponds to the inversion of the generative process. The model is divided into three parts that were implemented with different algorithms (see Methods). **Right:** cartoon (www.freepik.com) illustrating how a sequence of syllables 'lo-ri' (lorry) is generated from a traffic scene context. In describing a traffic accident, the speaker tries to convey its mental image of the scene consisting of an agent (the car), a patient (the lorry) and the relation (the action of hitting) from the agent to the patient. With English vocabulary and grammar, it chooses one lemma corresponding to each element in the accident, and outputs (speaks) these lemmas in a specific order according to the syntactic rules. Each lemma is then expressed as a specific sequence of syllables. Importantly, the same lemma can be the result of different combinations of abstract information and syntactic rules. For example, in the sentence "The ball hits the floor", the word "hits" implies a different action than car hitting a cyclist, whereas in "His songs are top hits" the relative position of the word implies an entity, not an action. **B. Temporal scheduling of hierarchical message passing during speech perception.** The generative model is inverted by alternating top-down prediction (prior, green downward arrows) and bottom-up update (blue upward arrows). A supraordinate level initiates a sequence of evidence accumulation in its subordinate level and receives a state update at the end of such sequence. It then sends an updated prediction to the subordinate level and initiates another sequence of evidence accumulation. Such process is repeatedly performed until the end of the sentence. Note that for the lemma and lower levels, states are generated anew each time when the supraordinate level makes a transition, i.e. no horizontal arrows between sending up an update and receiving a new prior. For the top two levels, however, states are maintained throughout the sequence (red horizontal arrows) or make transitions according to a set of rules (syntax).

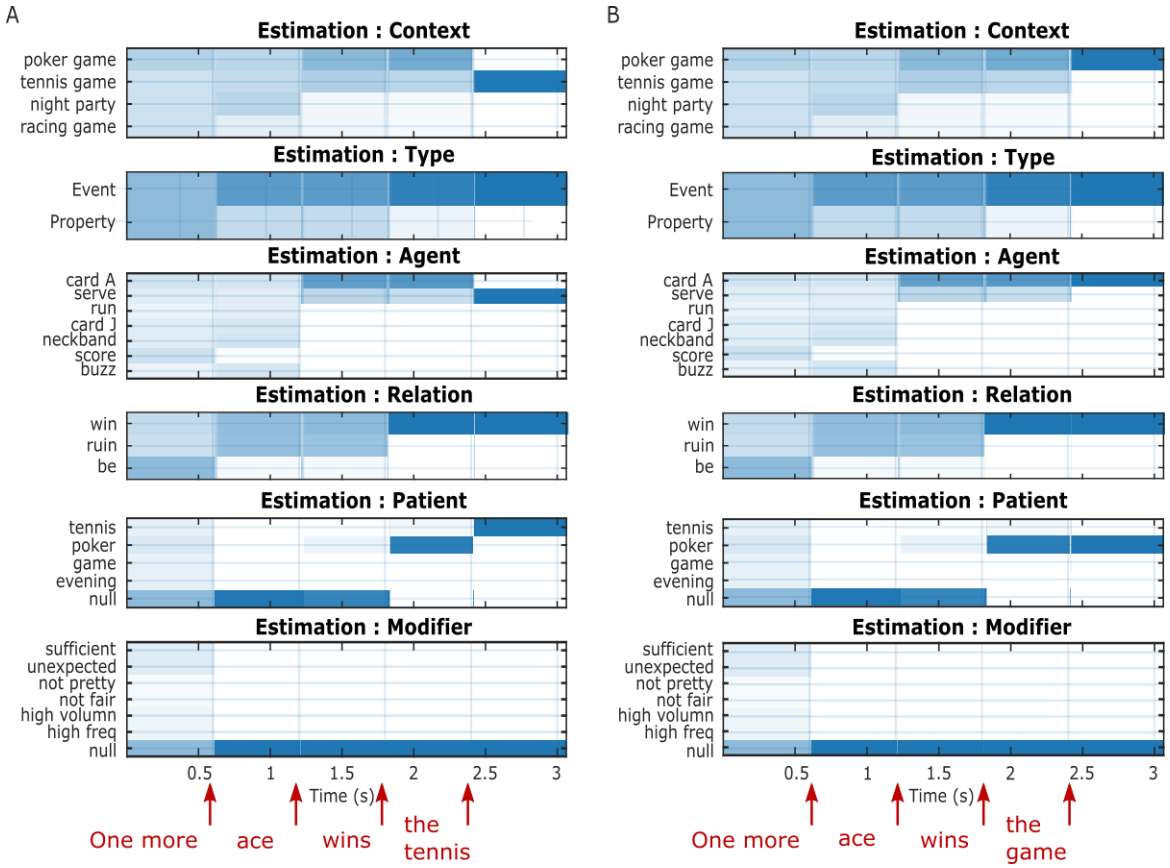

Figure 2. Model estimation of posterior probabilities for the semantic and context-level factors as sentences unfold. Possible values of each factor are labelled on the y axis. Blue scale blocks indicate the probability distribution for each factor, dark blue—p=1, white—p=0. Vertical lines and red arrows mark the offsets of lemma input, at which point updates were sent from the lemma level to semantic and context. The updating process is nearly instantaneous, and the main body of the n^{th} block (epoch corresponding to one lemma) is filled with the estimation after of the $(n-1)^{th}$ update. For both simulations, relative prior for context was set at the default of 1.5:1:1:1 for the four possibilities {'poker game', 'tennis game', 'night party', 'racing game'}. **A. Estimation for sentence A: "one more ace wins the tennis".** The first input "one more" was not informative. The estimated distributions were slightly changed before and after the offset of "one more" because the model still performed gradient descent to minimize free energy. After hearing "ace", distributions for the *context* and the *agent* converged to either "poker game" or "tennis game" for *context*, and 'card A' or 'serve' for *agent*. Within these possibilities, probabilities for the poker *context* and the 'card A' *agent* were higher, reflecting the prior preference. Probabilities of "tennis" or "poker" as *patient* also increased. *Type, relation*, and *modifier* remain the same as in the previous epoch. After hearing 'wins', possibilities for *type* converged to 'event', and those for *relation* converged to 'win'. Probabilities for 'tennis' and 'poker' for *patient* further increased, with a strong bias for "poker", while the probability of a 'null' *patient* decreased to zero. In the last epoch, the model received a disambiguating phrase 'the tennis', and all factors are resolved to the correct state with a probability close to 1. **B. Estimation for the sentence "one more ace wins the game".** The distributions are the same as in A before the last update. In the last epoch, the model receives an input, 'the game', that does not resolve the semantic and contextual ambiguity. As a result, distributions were further biased towards values corresponding to the 'poker game' context.

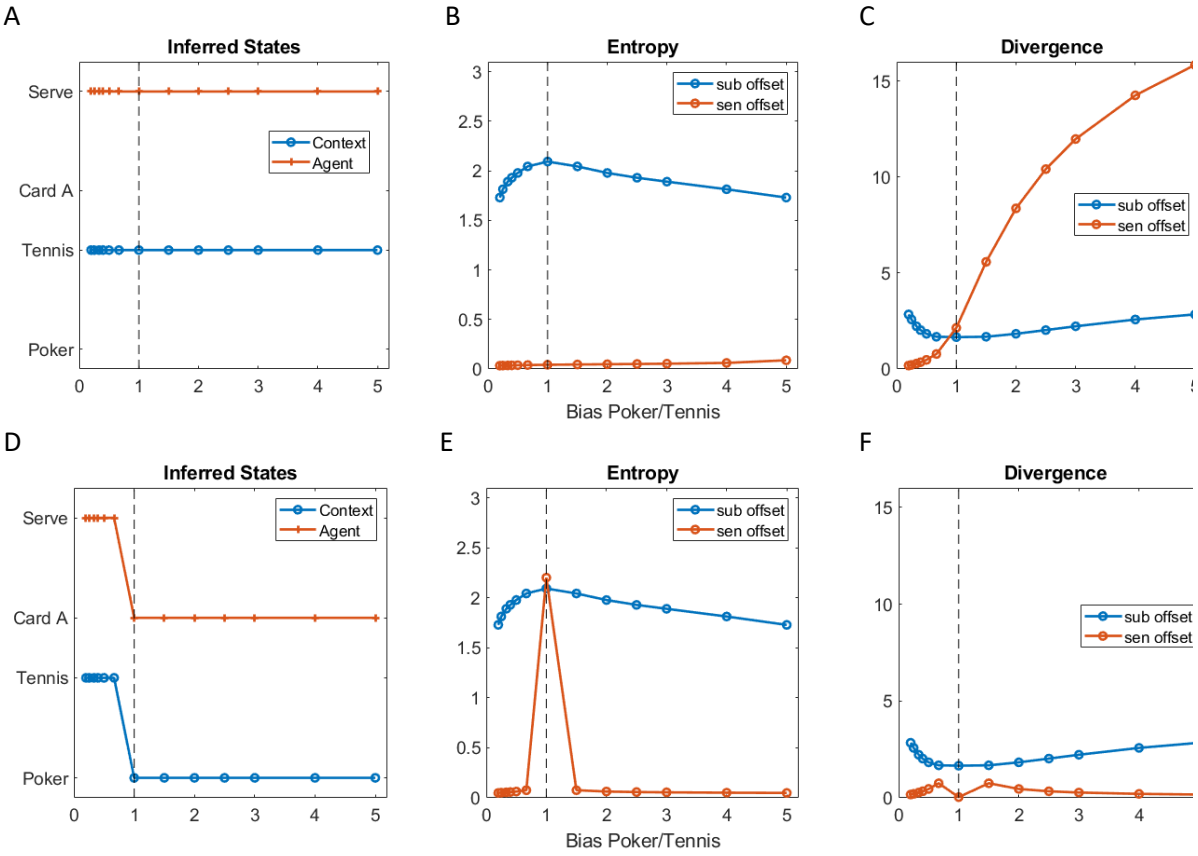

Figure 3. Effect of contextual bias ratio on the inference process. A-C: metrics derived from the sentence "One more ace wins the tennis" as function of contextual bias between "poker game" and "tennis game". A bias of x implies that the prior probability ratio (the total probability is always normalized to 1) for context was set to [x 1 1 1] for all 4 possible contexts ['poker game', 'tennis game', 'night party', 'racing game'} for x>=1, and [1 1/x 1 1] for x<1 to balance the influence of the two irrelevant contexts. **D-F: same metrics derived from sentence "One more ace wins the game". A.** Inferred states for the *context* (blue) and the *agent* (red) do not change with contextual bias, i.e. the model always resolved to the correct states. **B.** Sum of entropy across *context, agent* and *patient* at the subject word ("ace") offset and the sentence offset. At the offset of "ace" (blue), the entropy is maximum at bias=1 and symmetric on both sides. At sentence offset (red), the entropy is overall lower than at the offset of "ace" and monotonically increases with a small slope, reflecting that the model was more certain about the state estimations at this point, but keeps a small possibility towards the poker game that increases with the bias towards the poker context. **C.** At the sentence offset, the divergence monotonically increases with bias towards poker reflecting the increasing difference between the expected context (poker) and the actual one (tennis). **D.** Inferred states for context and agent at the end of sentence B as a function of bias. For bias<1 (preference for 'tennis'context), the inferred context is "tennis (game)" and inferred agent is "serve". For bias>=1, the result corresponds to a preference for the "poker" context. **E.** Sum of entropy. For both time points, the entropy is at maximum when bias=1. Both curves are symmetrical by bias=1. The blue curve is the same as in B because the sentence input up to this point was the same. **F.** Sum of divergence across the same three factors at two critical time points. At the offset of "ace", the divergence reached its minimum at bias=1 as a result of the uniform distribution over "poker" and "tennis" states, which is the least different from the previous time point. At the sentence offset, the stronger the bias (farther from 1), the smaller the difference between before and after hearing the final word. However, a notch is seen at bias=1 due to the uncertainty (Figure S1E).

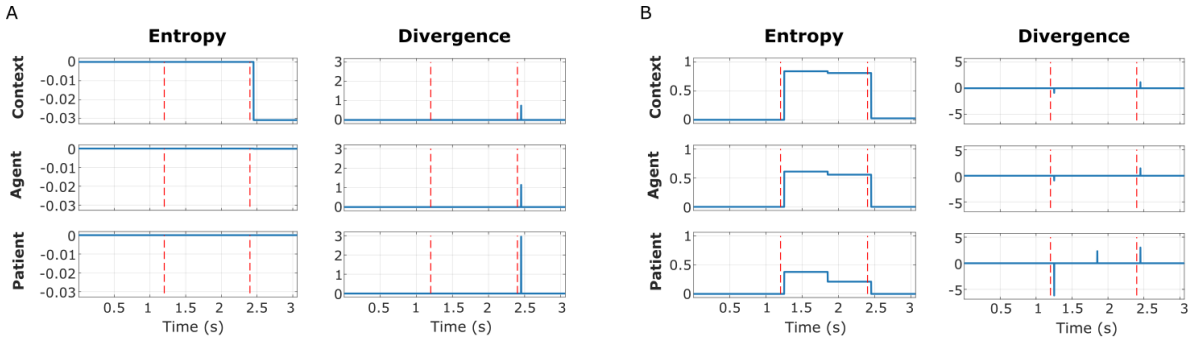

Figure 4. Differences of entropy and KL divergence for context, agent and patient between sentences. A. Entropy and divergence derived from sentence "one more ace wins the tennis" minus sentence "one more ace wins the game". The two vertical dashed lines mark the offset of the sentence middle word "ace" and the ending word, respectively. As the two sentences only differ in the ending word, both IT metrics differ only at the sentence offset. Compared to "the game", which does not completely resolve the ambiguity introduced by 'ace'", 'the tennis' results in lower entropy in "context" (top left panel), indicating greater certainty about the estimate. The zero differences in entropy for agent and patient indicate that the model tends to believe in its bias for these two factors. "The tennis" also gives rise to higher divergence (right panels) at sentence offset. **B. Results from sentence "one more ace wins the tennis" minus "one more sprint wins the tennis".** At its offset, the ambiguous word "ace" introduces higher entropy for all three factors compared to "sprint", reflecting the greater uncertainty about the hidden states. Such uncertainty dominates the divergence, which shows a corresponding negative difference here. At the sentence offset, differences of entropy between the two sentences became minimal because the model has resolved all factors. The positive difference in divergence at the offset reflects the higher surprisal for "the tennis" when it follows "ace" compared to "sprint".

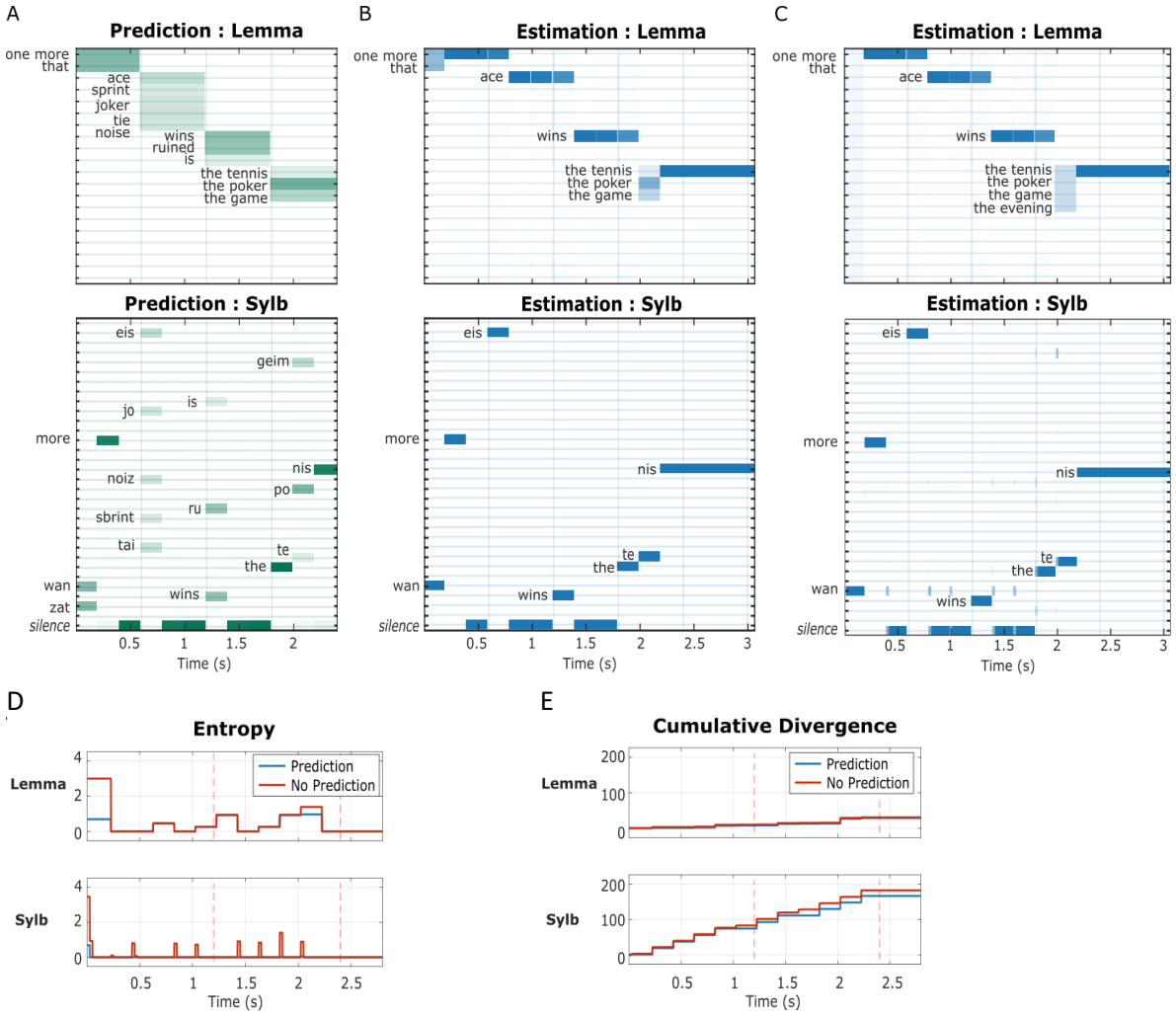

Figure 5. Influence of top-down predictions on syllable and lemma inference under high peripheral precision. All results are simulated with the sentence "One more ace wins the tennis". **A. Predictions (prior expectations sent from the superordinate level) for the simulation in Figure 2A.** Yellow vertical lines indicate offsets of each lemma input. In lemma 1-3, syllable predictions (lower panel) are nearly certain after the first syllable because there was a one-to-one correspondence between the lemma and the first syllable. In lemma 4 ("the tennis"), the opposite is true because all possible lemmas start with the syllable "the", diverging at the second syllable. Lemma predictions (top panel) depend on the current estimates at the superordinate level and the contextual bias, e.g. the prediction for the last lemma is highest for "the poker", and lowest for "the tennis". **B. Estimation of posterior probabilities for lemma and syllable states for the simulation in Figure 2A.** The model instantly recognizes each syllable (lower panel). The estimation for lemma states (upper panel) appears to lag for the duration of one syllable, because the lemma level receives a nearly instantaneous update at the **offset** of every syllable, and the grid between the i^{th} and $(i+1)^{th}$ updates is filled with the estimated distribution of the i^{th} update. For example, the estimation for the first lemma started with a 1:1 prior expectation between "one more" and "that", then converged to "one more" after hearing the first syllable "one". The estimation was not changed until the offset of "ace", the first syllable of the second lemma. This is only due to our illustrating method and does not affect the update from lemma to semantics. **C. Estimation of posterior probabilities when top-down predictions are set to uniform distributions for all possible states.** There are two noticeable differences at the syllable level (bottom panel) compared to B: 1) the initial estimate is a uniform prior for every syllable indicated by a very light grey vertical line through all possible states 2) a slight delay for the convergence of every syllable indicated by the small vertical bars, each corresponding to one spectral vector, in more than one possible state. The inference for lemma states is not significantly changed: once the model is certain about the first (or the second in the case of the last lemma) syllable, it can quickly converge to the correct lemma using its internal knowledge. **D. Contrast of entropy in the two prediction conditions calculated from lemma and syllable states.** With uninformative top-down prediction (red), the entropy of syllable states was raised for a short duration (~1-2 spectral vectors) more often than with informative prediction (blue). The difference is less obvious at the lemma level except during the very first syllable and the /the/ syllable in the last lemma. **E. Cumulative KL divergence in the two conditions.** Overall, the cumulative divergence is smaller when informative prediction is utilized (blue).

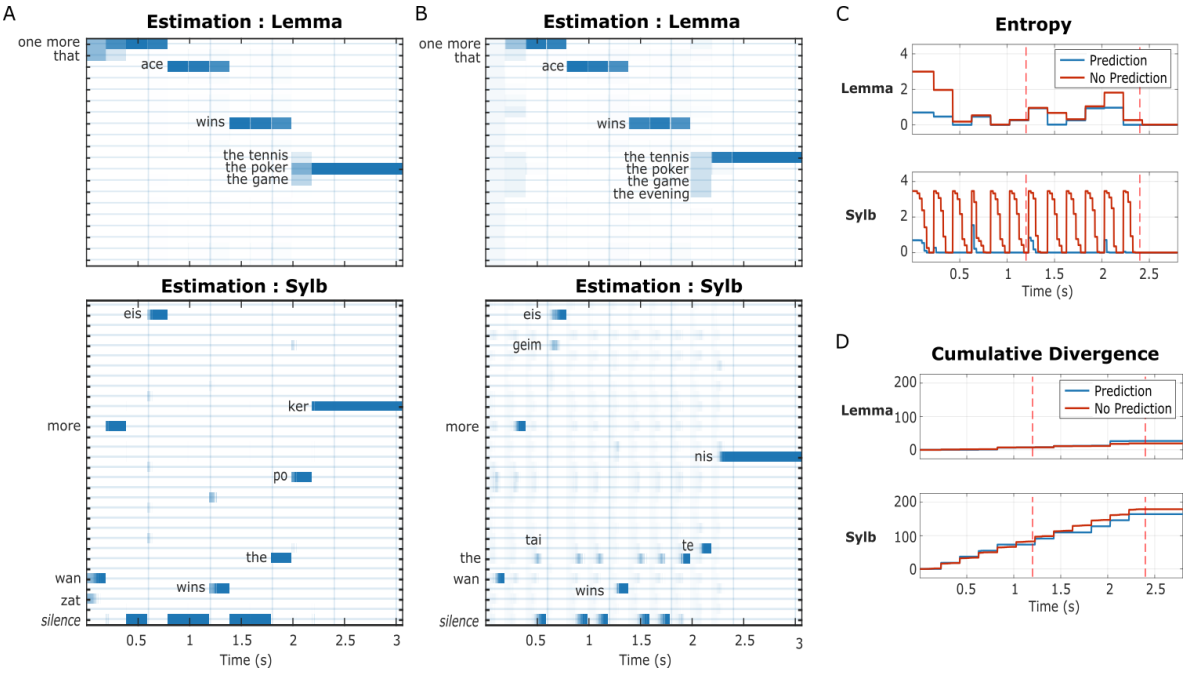

Figure 6. Influence of prediction with lowered peripheral precision. The input sentence, as in Figure 4 was "One more ace wins the tennis". Precision was set to p=exp(0), whereas in the intact model (Figure 4) p=exp(16). A and B: state estimation with and without informative prior. C and D: entropy and divergence in the two conditions. **A.** With informative prediction, the result is similar to that in Fig. 2A, except that 1) for the last lemma input, the model relied on the prediction, biased towards "the poker", and made the wrong inference, and 2) for the starting syllable in each lemma, the model took several spectral vectors to converge as indicated by the colored bars. **B.** Without prediction, the model took longer to infer each syllable compared to A, but inference was correct. **C. Entropy with informative (blue) or uninformative (red) top-down prediction for lemma and syllable estimates.** Without informative prediction, the uncertainty raised at the onset of every syllable instead of only for the syllable with multiple possible candidates (e.g. the syllable after "the" in the last lemma), and also reached higher magnitude as well as longer duration compared to the informative condition. **D. Cumulative divergence in the two conditions.** The divergence for syllable states was lower with informative prediction, but not for lemma. However, the summed divergence of the two levels is slightly higher with uninformative prediction.

Methods

We model speech perception by inverting a generative model of speech that is able to generate semantically meaningful sentences to express possible facts about the world. Since our main goal is to illustrate the cognitive aspect of speech comprehension, we use the model to simulate a semantic disambiguation task similar to MacGregor et al. (MacGregor et al., 2020). The task assesses the semantic ambiguity early in a sentence, which is disambiguated later in the sentence on half of the trials. Speech inputs to the model were synthesized short sentences adapted from MacGregor et al. (MacGregor et al., 2020).

In the next section we describe the speech stimuli, present the generative model, and briefly describe the approximate inversion of the generative model as well as the two information theoretic measures that could be related to measurable brain activity.

1. Speech stimuli

   In the original design of MacGregor and colleagues, eighty sentence sets were constructed to test the subjects' neural response to semantic ambiguity and disambiguation. Each set consists of four sentences in which two sentence MIDDLE WORDS crossed with two sentence final words. From the two sentence middle words, one was semantically ambiguous and from the two sentence final words one disambiguated the ambiguous middle word, and the other did not resolve the ambiguity. For example:

   The man knew that one more ACE might be enough to win the tennis.

   The woman hoped that one more SPRINT might be enough to win the game.

   The middle word was either semantically ambiguous ("ace" can be a special serve in a tennis game, or a poker card) or not ("sprint" only has one meaning of fast running); the two ending words either resolved the ambiguity of the middle word ("tennis" resolves "ace" to mean the special serve, not the poker card) or not ("game" can refer to either poker or tennis game). We chose this set as part of input stimuli to the model, but reduced the sentences to essential components for simplicity:

   One more ACE/SPRINT wins the tennis/game.

   The four sentences point to a minimum of two possible contexts, i.e. the nonlinguistic backgrounds where they might be generated: all combinations can result from a "tennis game" context, and the ACE-game combination can additionally result from a "poker game" context. Importantly, in our model the context is directly related to the interpretation of the word "ace".

   To balance the number of plausible sentences for each context, we added another possible mid-sentence word "joker", which unambiguously refers to a poker card in the model's knowledge. We also introduced another possible sentence structure to add syntactic variability within the same contexts:

   One more ACE/SPRINT is surprising/enough.

   The two syntactic structures correspond to two different types of a sentence: the "win" sentences describe an event, whereas the "is" sentences describe a property of the subject.

   We chose a total of two sentence sets from the original design. The other set (shortened version) is:

   That TIE/NOISE ruined the game/evening.

In these sentences, the subject "tie" can either mean a piece of cloth to wear around the neck ("neckband" in the model) or equal scores in a game. The ending word "game" resolves it to the latter meaning, whereas "evening" does not disambiguate between the two meanings. Similar to set 1, we added the possibility of property-type sentences. Table 1 lists all possible sentences and their corresponding contexts within the model's knowledge (ambiguous and resolving words are highlighted).

Table 1. All possible sentences in the model

Attribute	Subject	Verb	Object/Adjective	Context
One more/that	**ace**	wins	the game/**the poker**	poker game
			the game/the tennis	tennis game
		is	surprising/enough	poker or tennis game
	sprint	wins	the game/the tennis	tennis game
		is	surprising/enough	tennis game
	joker	wins	the game/the poker	poker game
		is	surprising/enough	poker game
	tie	ruined	the evening	night party/racing game
			the game	racing game
		is	**ugly**	**night party**
			unfair	**racing game**
	noise	ruined	the evening/the game	night party/racing game
		is	loud/sharp	night party/racing game

The input to the model consisted of acoustic spectrograms that were created using the Praat speech synthesizer (Boersma, 2021) with British accent, male speaker 1.

In this work we are not focusing on timing or parsing aspects, rather on how information is incorporated into the inference process in an incremental manner and how the model's estimates about a preceding word can be revised upon new evidence during speech processing. Therefore, we chose the syllable as the interface unit between continuous and symbolic representations, and fixed the length of the input to simplify the model construction (see details in Generative model). Each sentence consists of four lemma items (single words or two-word phrases), and each lemma consists of three syllables. All syllables were normalized in length by reducing the acoustic signal to 200 samples.

Specifically, in Praat, we first synthesized full words, then separated out syllables using the TextGrid function. A 6-by-200 time-frequency (TF) matrix was created for each unique syllable by averaging its spectro-temporal pattern into 6 log-spaced frequency channels (roughly spanning from 150 Hz to 5 kHz) and 200 time bins in the same fashion as in Hovsepyan et al. (Hovsepyan et al., 2020). Each sentence input to the model was then assembled by concatenating these TF matrices in the appropriate order. Since we fixed the number of syllables in each word (Ns = 3), words consisting of fewer syllables were padded with "silence" syllables, i.e. all-zero matrices. During simulation, input was provided online in that 6-by-1 vectors from the padded TF matrix representing the full sentence were presented to the model one after another, at the rate of 1000 Hz. In effect, all syllables were normalized to the same duration of 200ms. The same TF

matrices were used for the construction of the generative model as speech templates (see section 2c for details).

2. Generative model

The generative model goes from a nonlinguistic, abstract representation of a message defined in terms of semantic roles to a linearized linguistic sentence and its corresponding sound spectrogram. The main idea of the model is that listeners have knowledge about the world that explains how an utterance may be generated to express a message from a speaker.

In this miniature world, the modeled listener knows about a number of *contexts*, the scenarios under which a message is generated (to distinguish them from names given to representation levels in the model, we will use *italic* to refer to factors at each level; see below). Each message can either be of an "event" *type* that describes an action within the context, or of a "property" *type* that expresses a characteristic of an entity that exists in the context. *Context* and *type* are nonlinguistic representations maintained throughout the message but make contact with linguistic entities via semantics and syntax, which jointly determine an ordered sequence of lemma that then generates the acoustic signal of an utterance that evolves over time.

As in the real world, connections from context to semantics and semantics to lemma are not one-to-one, and ambiguity arises, for example, when two semantic items can be expressed as the same lemma. In this case the model can output exactly the same utterance for two different messages. When the model encounters such an ambiguous sentence during inference, it will make its best guess based on its knowledge when ambiguity is present (see Model inversion). For illustrative purposes, we only consider a minimum number of alternatives, sufficient to create ambiguity, e.g. the word "ace" only has two possible meanings in the model. Also, while the model generates a finite set of possible sentences, they are obtained in a compositional fashion; they are not spelled out explicitly anywhere in the model, and must be incrementally constructed according to the listener's knowledge.

Specifically, the generative model (Figure 1A) is organized in three hierarchically related submodels that differ in their temporal organization, with each submodel providing empirical priors to the subordinate submodel, which then evolves in time according to its discrete or continuous dynamics for a fixed duration (as detailed below). Overall, this organization results in six hierarchically related levels of information carried by a speech utterance, from high to low (L_1-L_6) we refer to them as: context, semantics and syntax, lemma, syllable, acoustic, and the continuous signal represented by time-frequency (TF) patterns that stands for the speech output signal.

Each level in the model consists of one or more factors representing the quantities of interest (e.g., *context, lemma, syllable …*), illustrated as rectangles in Fig 1A. We use the term "states" or hidden states to refer to the values that a factor can take (e.g. in the model the factor *context* can be in one of four states ['poker game', 'tennis game', 'night party', 'racing game'}. For a complete list of factors and their possible states of context to lemma levels see Table 2).

Table 2. Factors and their possible values (states) in the model hierarchy

Level	Factor	Value
1	Context	Tennis game, poker game, night out, car racing game
1	Sentence type	Event, property
2	Agent (semantic)	Card A, winning serve, run, card J, neckband, score, buzz, null
2	Patient (semantic)	Tennis game, poker game, racing game, evening, null
2	Relation (semantic)	Win, ruin, be

2	Modifier (semantic)	Sufficient, unexpected, not pretty, not fair, high volumn, high frequency
2	Syntax	Attribute, subject, verb, object, adjective
3	Lemma	One more, that, ace, sprint, joker, tie, noise, wins, ruined, is, the tennis, the poker, the game, the evening, enough, surprising, ugly, unfair, loud, sharp
3	Where in lemma	1-3
4	Syllable*	/eis/, /te/, /nis/, … total of 32 including the silence syllable
4	Where in syllable	1-8

*Note that these symbols are illustrative and not strictly following IPA.

As an example, to generate a sentence to describe an event under a "tennis game" *context*, the model picks "tennis serve" as the agent, "tennis game" as the patient, and "win" as their relationship. When the syntactic rule indicates that the current semantic role to be expressed should be the agent, the model selects the lemma "ace", which is then sequentially decomposed into three syllables /eis/, /silence/, /silence/. Each syllable corresponds to eight 6-by-1 spectral vectors that are deployed in time over a period of 25 ms each. The generative model therefore generates the output of continuous TF patterns as a sequence of "chunks" of 25 ms.

We next describe in detail the three submodels:

a. Discrete non-nested: context to lemma via semantic (dependency) and syntax (linearization)

The context level consists of two independent factors: the *context c* and the sentence *type Ty*. Together, they determine the probability distribution of four semantic roles: the *agent* s^A, the *relation* s^R, the *patient* s^P, and the *modifier* s^M. An important assumption of the model is that states of *context*, *type* and semantic roles are maintained throughout the sentence as if they had memory. These semantic roles generate a sequence of lemmas in the subordinate level, whose order is determined by the *syntax,* itself determined by the sentence *type*. This generative model for the first to the n[th] lemma is ($\vec{s}$ denotes the collection of all semantic factors $\vec{s} = [s^A, s^R, s^P, s^M]$):

$$p(w^1, \cdots, w^n, syn^1, \cdots, syn^n, \vec{s}, c, Ty) = p(w^1|syn^1, \vec{s}) \cdots p(w^n|syn^n, \vec{s}) p(\vec{s}|c, Ty) p(c) p(syn^1, \cdots, syn^n|Ty) p(Ty) \quad (1)$$

Here, p(c) is the prior distribution for the *context*. The prior probability for the sentence type p(Ty) was fixed to be equal between "property" and "event".

The terms $p(\vec{s}|c, Ty)$ and $p(syn^1, \cdots, syn^n|Ty)$ can be further expanded as:

$$p(\vec{s}|c, Ty) = p(s^A|c) p(s^R|c, Ty) p(s^P|c, Ty) p(s^M|c, Ty) \quad (2)$$

$$p(syn^1, \cdots, syn^n|Ty) = p(syn^1|Ty) \cdots p(syn^n|Ty) \quad (3)$$

When *Ty*='event', the sentence consists of an *agent*, a *patient*, a *relation* between the *agent* and the *patient*, and a null (empty) *modifier*. When *Ty*='property', the sentence consists of an *agent*, a *modifier* that describes the *agent*, a *relation* that links the *agent* and the *modifier*, and a null *patient*.

To translate the static context, type and semantic states into ordered lemma sequences, we constructed a minimal (linear) syntax model consistent with English grammar. We constrain all possible sentences to have four syntactic elements syn^1-syn^4, values are ['attribute', 'subject', 'verb', 'object', 'adjective'}. The probability of syn^n is dependent solely on Ty.

The syntactic element syn^i is active during the i[th] epoch, and each possible value of the syntax (except 'attribute' that directly translates to a lemma item randomly determined

within ['one more' and 'that'}) corresponds to one semantic factor (semantic factors in the model include subject, verb, object and adjective):

Subject—*agent ;* Verb—*relation ;* Object—*patient ;* Adjective—*modifier*

Thus, sentences of the "event" type are always expressed in the form of subject-verb-object (SVO), and those of the "property" type in the form of subject-verb-adjective (SVadj). In the i^{th} lemma epoch, the model picks the current semantic factor via the value of syn_i and finds a lemma to express the value (state) of this semantic factor, using its internal knowledge of mapping between abstract, nonlinguistic concepts to lexical items (summarized in the form of a dictionary in Appendix I). Note that the same meaning can be expressed by more than one possible lemma, and several different meanings can result in the same lemma, causing ambiguity. The mapping from L_2 to L_3 can be defined separately for each lemma as follows:

- The first lemma (w^1 the attribute) does not depend on semantics or syntax and the model would generate "one more" or "that" with equal probability (p=0.5).
- w^2 and w^3 are selected according to *agent* and *patient* values, respectively, which are themselves constrained by context.
- w^4 can be either a patient or a modifier depending on Ty.

Prior probabilities of context and type, as well as probabilistic mappings between levels (eq.2-4), are all defined in the form of multidimensional arrays. Detailed expressions and default values can be found in Appendix II.

b. Discrete nested: lemma to spectral

Over time, factors periodically make probabilistic transitions between states (not necessarily different). Different model levels are connected in that during the generative process, discrete hidden (true) states of factors in a superordinate level (L_n) determine the initial state of one or more factors in the subordinate level (L_{n+1}). The L_{n+1} factors then make a fixed number of state transitions. When the L_{n+1} sequence is finished, L_n makes one state transition and initiates a new sequence at L_{n+1}. State transitioning of different factors within the same level occurs at the same rate. We refer to the time between two transitions within each level as one **epoch** of the level. Thus, model hierarchies are temporally organized in that lower levels evolve at higher rates and are nested within their superordinate levels.

The formal definition of the discrete generative model is shown in eq.1, where the joint probability distribution of the m^{th} outcome modality (here generally denoted by o^m, specified in following sections) and hidden states (generally denoted by s^n) of the n^{th} factor up to a time point τ, is determined by the priors over hidden states at the initial epoch $P(s^{n,1})$, the likelihood mapping from states to outcome $P(o|s)$ over time 1:τ, and the transition probabilities between hidden states of two consecutive time points $P(s^{n,t}|s^{n,t-1})$ up to t=τ:

$$P(o^{m,1:\tau}, s^{n,1:\tau}) = P(s^{n,1}) \prod_{\tau} P(o^{m,\tau}|s^{n,\tau}) P(s^{n,\tau}|s^{n,\tau-1}) \qquad (4)$$

For lower discrete levels, representational units unfold linearly in time, and a sequence of subordinate units can be entirely embedded within the duration of one superordinate epoch. Therefore, the corresponding models are implemented in a uniform way: the hidden state consists of a "what" factor that indicates the value of the representation unit (e.g. the lemma 'the tennis'), and a "where" factor that points to the location of the outcome

(syllable) within the "what" state (e.g. the 2nd location of 'tennis' generates syllable '/nis/'). During one epoch at each level (e.g. the entire duration of the lemma "the tennis"), the value of the "what" factor remains unchanged with its transition probabilities set to the unit matrix. The "where" factor transitions from 1 to the length of the "what" factor, which is the number of its subordinate units during one epoch (three syllables per lemma). Together, the "what" and "where" states at the lemma level generate a sequence of syllables by determining the prior for "what" and "where" states in each syllable. In the same fashion, each syllable determines the prior for each spectral vector. Thus, the syllable level goes through 8 epochs, and for each epoch the output of the syllable level corresponds to a spectral vector of dimension (1 x 6, number of frequency channels). This single vector determines the prior for the continuous submodel.

Such temporal hierarchy is roughly represented in Figure 1B (downward arrows).

Unlike L_1 and L_2 states that are maintained throughout the sentence, states of the lemma level and below are "memoryless", in that they are generated anew by superordinate states at the beginning of each epoch. This allows us to simplify the model inversion (see next section) using a well-established framework that exploits the variational Bayes algorithm for model inversion (Friston et al., 2017). The framework of Friston et al. (Friston et al., 2017) consists of two parts: hidden state estimation and action selection. In our model, the listener does not perform any overt action (the state estimates do not affect state transitioning), therefore the action selection part is omitted.

Using the notation of Eq.1, parameters of the generative model are defined in the form of multidimensional arrays:

Probabilistic mapping from hidden states to outcomes: $P(o^{m,\tau}|s^{1,\tau},...,s^{N,\tau}) = Cat(A^m)$ (5)

Probabilistic transition among hidden states: $P(s^{n,\tau+1}|s^{n,\tau}) = Cat(B^{n,\tau})$ (6)

Prior beliefs about the initial hidden states: $P(s^{n,1}) = Cat(D^n)$ (7)

For each level we define **A, B, D** matrices according to the above description of hierarchical "what" and "where" factors:

- Probability mappings (matrix **A**) from a superordinate "what" to a subordinate "what" states are deterministic, e.g. p(sylb='/one/'|lemma='one more', where=1)=1, and no mapping is needed for "where" states;
- Transition matrices (**B**) for "what" factors are all identity matrices, indicating that the hidden state does not change within single epochs of the superordinate level;
- Transition matrices for "where" factors are off-diagonal identity matrices, allowing transition from one position to the next;
- Initial states (**D**) for "what" factors are set by the supraordinate level, and always start at position 1 for "where" factors.

c. Continuous: acoustic to output

The addition of an acoustic level between the syllable and the continuous levels is based on a recent biophysically plausible model of syllable recognition, Precoss (Hovsepyan et al., 2020). In that model syllables were encoded with continuous variables and represented, as is the case here, by an ordered sequence of 8 spectral vectors (each vector having six components corresponding to six frequency channels). In the current model we only implemented the bottom level of the Precoss model (see also (Yildiz et al., 2013)), which

deploys spectral vectors into continuous temporal patterns. Specifically, the outcome of the syllable level sets the prior over the hidden cause, a spectral vector **I** that drives the continuous model. It represents a chunk of the time-frequency pattern determined by the "what" and "where" states of the syllable level s^{ω} and s^{γ} respectively:

$$I_f = \sum_{\omega=1}^{Nsyl} \sum_{\gamma=1}^{8} s^{\omega} s^{\gamma} V_{f\omega\gamma} + \epsilon^I \quad (8)$$

$$V_{f\omega\gamma} = G_f(TF_{\omega\gamma}) - W_f \tanh(TF_{\omega\gamma}) \quad (9)$$

The noise terms ε^{I} is random Gaussian fluctuation. $TF_{\omega\gamma}$ stands for the average of the 6x200 TF matrix of syllable ω in the γ^{th} window of 25 ms. **G** and **W** are 6x6 connectivity matrices that ensure the spectral vector **I** determines a global attractor of the Hopfield network that sets the dynamics of the 6 frequency channels. Values of **G**, **W** and a scalar rate constant κ in eq. 9-10 are the same as in Precoss:

$$\frac{dx}{dt} = \kappa[-Gx + W \tanh x + I] + \epsilon^x \quad (10)$$

The continuous state of **x** determines the final output of the generative model **v**, which is compared to the speech input during model inversion. As **x, v**, is a 6x1 vector:

$$v = x + \epsilon^v \quad (11)$$

The precision of the output signal depends on the magnitude of the random fluctuations in the model (ε in eq. 8, 10, 11). During model inversion, the discrepancy between the input and the prediction of the generative model, i.e. the prediction error, are weighted by the corresponding precisions and used to update model estimates in generalized coordinates (Friston et al., 2008). We manipulated the precisions for continuous state **x** and activities of frequency channels **v** to simulate intact (HP) and impaired (LP) peripheries. The precision for top-down priors from the syllable level, Ps, was kept high for all simulations (see Table 2 for values used in different conditions).

The continuous generative model and its inversion were implemented using the ADEM routine in the SPM12 software package (Neuroimaging, 2014), which integrates a generative process of action. Because we focus on passive listening rather than interacting with the external world, this generative process was set to identical to the generative model and without an action variable. Precisions for the generative process were the same for all simulations (Table 3).

Table 3. Precisions

Precision	Generative model: HP	Generative model: LP	Generative process
$\mathbf{P^x}$	exp(16)	exp(0)	exp(16)
$\mathbf{P^v}$	exp(16)	exp(0)	exp(16)
$\mathbf{P^I}$	exp(8)	exp(8)	exp(8)

3. Model inversion

The goal of the modeled listener is to estimate posterior probabilities of all hidden states given observed evidence p(s|o), which is the speech input to the model, here represented by TF patterns sampled at 1000 Hz. This is achieved by the inversion of the above generative model using the variational Bayesian approximation under the principle of minimizing free energy (Friston et al., 2006). Although this same computational principle is applied throughout all model hierarchies, the implementation is divided into three parts corresponding to the division of the generative model. Because the three "submodels" are hierarchically related we follow and adapt the approach proposed in (Friston et al., 2017), which shows how to invert models with

hierarchically related components through Bayesian model averaging. The variational Bayes approximation for each of the three submodels is detailed below.

Overall, the scheme results in a nested estimation process (Figure 1B). For a discrete-state level L_n, probability distributions over possible states within each factor are estimated at discrete times over multiple inference epochs. Each epoch at level L_n starts as the estimated L_n states generate predictions for initial states in the subordinate level L_{n+1}, and ends after a fixed number of state transitions (epochs) at L_{n+1}. State estimations for L_n are then updated using the discrepancy between the predicted and observed L_{n+1} states. The L_n factors make transitions into the next epoch immediately following the update, and the same process is repeated with the updated estimation. Different model hierarchies (from L_2 on) are nested in that the observed L_{n+1} states are state estimations integrating information from L_{n+2} with the same alternating prediction-update paradigm, but in a faster timescale. A schematic of such a hierarchical prediction-update process is illustrated in Figure 1B.

Since levels "lemma" to the continuous acoustic output conform to the class of generative models considered in (Friston et al., 2017), we use their derived gradient descent equations and implementation. Levels "context" and "semantic and syntax" do not conform to the same class of discrete models (due to their memory component and non-nested temporal characteristics); we therefore derived the corresponding gradient descent equations based on free energy minimization for our specific model of the top two levels Equations 2-4 (see Appendix III for the derivation) and incorporated them into the general framework of (Friston et al., 2017).

The variational Bayes approximation for each of the three submodels is detailed below.

a. Lemma to context

For all discrete-state levels, the free energy F is generally defined as: (Friston et al., 2006).

$$Q(s) = arg\min_{Q(s)} F \approx P(s|o) \quad (12)$$

$$F = E_Q[\ln Q(s) - \ln P(o|s) - \ln P(s)] \quad (13)$$

In eq. 12 and 13, Q(s) denotes the estimated posterior probability of hidden state s, P(o|s) the likelihood mapping defined in the generative model, and P(s) the prior probability of s. The variational equations to find the Q(s) that minimizes Free energy can be solved via gradient descent. We limit the number of gradient descent iterations to 16 in each update to reflect the time constraint in neuronal processes.

Although context/type and semantic/syntax are modeled as two hierarchies, we assign them the same temporal scheme for the prediction-update process at the rate of lemma units, i.e. they both generate top-down predictions prior to each new lemma input, and fulfill bottom-up updates at each lemma offset. Therefore, it is convenient to define their inference process in conjunction.

The posterior distribution $p(syn^1, \cdots, syn^n, \vec{s}, c, ST|w^1, \cdots, w^n)$ is approximated by a factorized one, $Q(syn^1)\cdots Q(syn^n)Q(s^1)\cdots Q(s^{n_s})Q(c)Q(ST)$, and is parameterized as follows:

$$
\begin{aligned}
Q(syn^{\tau}) &: \quad syn_k^{(\tau)}, \quad \text{or} \quad \mathcal{C}at(syn^{(\tau)}), \quad k = 1, \cdots, \#\text{ of possible syntactic elements} \\
&\quad\ \tau = 1, \cdots, n \\
Q(s^{\alpha}) &: \quad s_j^{(\alpha)}, \quad \text{or} \quad \mathcal{C}at(s^{(\alpha)}) \quad j = 1, \cdots, \#\text{ of possible states for semantic factor } \alpha \\
&\quad\ \alpha = \{A, R, P, M\} \\
Q(c) &: \quad c_m, \quad \text{or} \quad \mathcal{C}at(c) \quad m = 1, \cdots, \#\text{ of possible states for context factor} \\
Q(Ty) &: \quad Ty_a \quad \text{or} \quad \mathcal{C}at(Ty) \quad a = 1, \cdots, \#\text{ of possible states for sentence type}
\end{aligned}
$$

Here, the model observation is the probability of the word being w^{τ} given the observed outcome o^{τ}, $p(w^{\tau}|\, o^{\tau})$, which is gathered from lower-level models described in next sections. We denote $p(w^{\tau}|\, o^{\tau})$ by a vector W_i^{τ}, where τ stands for the epoch, and *i* indexes the word in the dictionary. At the beginning of the sentence, the model predicts the first lemma input, which is, by definition, just one of the two possible attributes, 'one more' or 'that'.

$$
\begin{aligned}
p(w^1) &= \sum_{syn^1, \vec{s}, c, Ty} p(w^1|syn^1, \vec{s}, c, Ty) p(syn^1, \vec{s}, c, Ty) \\
&= \sum_{syn^1} p(w^1|syn^1) p(syn^1) = p(w^1|syn^1 = attribute) \qquad (14)
\end{aligned}
$$

The lower levels then calculate $p(w^1|o^1)$ and provide an updated W_i^1 that incorporates the observation made from the first lemma. This is passed to the top levels to update L_1 and L_2 states. Following this update, the next epoch is initiated with the prediction for w^2. Because w^2 does not directly depend on lemma inputs before and after itself, we can derive the following informed prediction of w^2 from eq.2, where prior for L_1 and L_2 factors are replaced by their updated posterior estimates:

$$
\begin{aligned}
p(w^2) &= \sum_{syn^2, \vec{s}, c, ST} p(w^2|syn^2, \vec{s}, c, Ty) p(syn^2, \vec{s}, c, Ty|o^1) \\
&\approx \sum_{syn^2, \vec{s}, Ty} p(w^2|syn^2, \vec{s}) p(syn^2|Ty) Q^{(1)}(\vec{s}) Q^{(1)}(c) Q^{(1)}(Ty) \qquad (15)
\end{aligned}
$$

Where we used:

$$
\begin{aligned}
p(syn^2, \vec{s}, c, Ty|o^1) &\approx p(syn^2|Ty) Q(\vec{s}, c, Ty|o^1) \\
&= p(syn^2|Ty) Q^{(1)}(\vec{s}) Q^{(1)}(c) Q^{(1)}(Ty)
\end{aligned}
$$

During the second epoch, the model receives input of the second lemma and updates the estimation of W_i^2. The updated W_i^2 is then exploited to update L_1 and L_2 states, which in turn provides the prediction for w^3. The process is repeated until the end of the sentence.

The updating of L_1 and L_2 states, i.e. the estimation of their posterior probabilities after receiving the n^{th} lemma input relies on the minimization of the total free energy $F_{1,2}$ of the two levels (L_1, L_2)

$$
\begin{aligned}
F_{1,2} \equiv \sum_{syn^1:syn^n, \vec{s}, c, Ty} & Q(syn^1, \cdots, syn^n, \vec{s}, c, Ty) \Big[\ln Q(syn^1, \cdots, syn^n, \vec{s}, c, Ty) \\
& - \sum_{w^1:w^n} Q(w^1, \cdots, w^n) \ln p(w^1, \cdots, w^n, syn^1, \cdots, syn^n, \vec{s}, c, Ty) \Big] \qquad (16)
\end{aligned}
$$

The expanded expression of $F_{1,2}$ and derivation of the gradient descent equations can be found in Appendix III.

b. Spectral to lemma

The memoryless property of lower-level (lemma and below) states implies that the observation from the previous epoch does not directly affect the prediction for the new epoch, only indirectly through the evidence accumulated at superordinate levels. The framework from Friston et al. (Friston et al., 2017) is suitable for such construction. It uses the same algorithm of free-energy (inserting eq. 5-7 to eq. 12-13) minimization for posterior estimation, but this time there is conditional independence between factors in the same level. We implemented this part of the model by adapting the variational Bayesian routine in the DEM toolbox from the SPM12 software package (Neuroimaging, 2014).

c. Continuous to spectral

To enable the information exchange between the continuous and higher discrete levels that were not accounted for in Hovsepyan et al. (Hovsepyan et al., 2020), we implemented the inversion of the spectral-to-continuous generative model using the "mixed model" framework in Friston et al. (Friston et al., 2017). Essentially, the dynamics of spectral fluctuation determined by each spectral vector **I** (eq.8) is treated as a separate model of continuous trajectories, and the posterior estimation of **I** constitutes post-hoc model comparison that minimizes free energy in the continuous format. For a specific model m represented by spectral vector I_m, the free energy $F(t)_m$ can be computed as (adapted from (Friston et al., 2017)):

$$F(t)_m = -\ln P(o_m) - \int_0^T L(t)_m dt \quad (17)$$

$$L(t)_m = \ln P(o(t)|I_m) - \ln P(o(t)|I) \quad (18)$$

$P(o_m)$ indicates the likelihood for the m^{th} spectral vector (discrete). $P(o(t)|I_m)$ is the likelihood of observing the continuous input o(t) given the m^{th} **I** vector, and $P(o(t)|I)$ is the averaged likelihood over all possible **I** vectors. In this way, the model compares the top-down prediction of I and the estimate derived from the bottom-up evidence of integrated acoustic input over 25ms. Detailed explanation of the algorithm can be found in previous studies (Friston et al., 2017, Friston and Penny, 2011). The software implementation was also adapted from existing routines in the DEM toolbox of SPM12 (Neuroimaging, 2014).

Information theoretic metrics

Two metrics were derived from the belief updating process just described: the Kullback-Leibler (KL) divergence (Div), which characterizes the discrepancy between the current and previous state estimates of a factor, and entropy H that characterizes the uncertainty of the current state estimates of the factor. We denote the posterior probability of the i^{th} possible state of an arbitrary factor at time point τ as q^t_i. The divergence and entropy are defined as:

$$Div^\tau = -\sum_i q_i^\tau \ln q_i^{\tau-1} + \sum_i q_i^\tau \ln q_i^\tau \quad (19)$$

$$H^\tau = -\sum_i q_i^\tau \ln q_i^\tau \quad (20)$$

These two (non-orthogonal) metrics provide a qualitative summary of the model response that can be linked to neurophysiological signals (see Result and Discussion).

Acknowledgements

We thank B. Bickel, S. van Ommen, D. Poeppel for critical feedback, and E. Holmes for advice on the SPM software. This work was funded by Swiss National Science Foundation (grant number 320030B_182855) and NCCR Evolving Language, Swiss National Science Foundation Agreement #51NF40_180888.

Data and code availability

Custom code and simulation data will be made available upon request.

Conflict of interest

The authors declare no conflict of interest.

References

ALTMANN, G. T. M. 1999. Thematic role assignment in context. *Journal of Memory and Language,* 41**,** 124-145.

ALTMANN, G. T. M. & MIRKOVIC, J. 2009. Incrementality and Prediction in Human Sentence Processing. *Cognitive Science,* 33**,** 583-609.

ARNAL, L. H. & GIRAUD, A. L. 2012. Cortical oscillations and sensory predictions. *Trends Cogn Sci,* 16**,** 390-8.

ASSMANN, P. & SUMMERFIELD, Q. 2004. The perception of speech under adverse conditions. *Speech processing in the auditory system.* Springer.

BARRETT, L. F. & SIMMONS, W. K. 2015. Interoceptive predictions in the brain. *Nature Reviews Neuroscience,* 16**,** 419-429.

BASTOS, A. M., LUNDQVIST, M., WAITE, A. S., KOPELL, N. & MILLER, E. K. 2020. Layer and rhythm specificity for predictive routing. *Proc Natl Acad Sci U S A,* 117**,** 31459-31469.

BASTOS, A. M., USREY, W. M., ADAMS, R. A., MANGUN, G. R., FRIES, P. & FRISTON, K. J. 2012. Canonical Microcircuits for Predictive Coding. *Neuron,* 76**,** 695-711.

BECK, J., HELLER, K. & POUGET, A. 2012. Complex inference in neural circuits with probabilistic population codes and topic models.

BENDER, E. M. & KOLLER, A. Climbing towards NLU: On meaning, form, and understanding in the age of data. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020. 5185-5198.

BLEI, D. M., GRIFFITHS, T. L., JORDAN, M. I. & TENENBAUM, J. B. Hierarchical topic models and the nested Chinese restaurant process. NIPS, 2003.

BOERSMA, P. W., DAVID 2021. Praat: doing phonetics by computer.

BROWN, T. B., MANN, B., RYDER, N., SUBBIAH, M., KAPLAN, J., DHARIWAL, P., NEELAKANTAN, A., SHYAM, P., SASTRY, G. & ASKELL, A. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

CASTELLUCCI, G. A., KOVACH, C. K., HOWARD, M. A., GREENLEE, J. D. W. & LONG, M. A. 2022. A speech planning network for interactive language use. *Nature*.

CAUCHETEUX, C. & KING, J. R. 2022. Brains and algorithms partially converge in natural language processing. *Communications Biology,* 5.

CHRISTIANSEN, M. H. & CHATER, N. 2016. The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences,* 39.

CLARK, A. 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences,* 36**,** 181-204.

CORBALLIS, M. C. 2009. The Evolution of Language. *Year in Cognitive Neuroscience 2009,* 1156**,** 19-43.

DELONG, K. A., URBACH, T. P. & KUTAS, M. 2005. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nat Neurosci,* 8**,** 1117-21.

DEMBERG, V. & KELLER, F. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition,* 109**,** 193-210.

DEVLIN, J., CHANG, M.-W., LEE, K. & TOUTANOVA, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

DING, N., MELLONI, L., ZHANG, H., TIAN, X. & POEPPEL, D. 2016. Cortical tracking of hierarchical linguistic structures in connected speech. *Nat Neurosci,* 19**,** 158-64.

DONHAUSER, P. W. & BAILLET, S. 2020. Two Distinct Neural Timescales for Predictive Speech Processing. *Neuron,* 105**,** 385-393 e9.

ECKERT, M. A., TEUBNER-RHODES, S. & VADEN, K. I. 2016. Is Listening in Noise Worth It? The Neurobiology of Speech Recognition in Challenging Listening Conditions. *Ear and Hearing,* 37**,** 101s-110s.

EGOROVA, N., SHTYROV, Y. & PULVERMULLER, F. 2013. Early and parallel processing of pragmatic and semantic information in speech acts: neurophysiological evidence. *Frontiers in Human Neuroscience,* 7.

ELMAN, J. L. 1990. Finding Structure in Time. *Cognitive Science,* 14**,** 179-211.

FAIRS, A., MICHELAS, A., DUFOUR, S. & STRIJKERS, K. 2021. The Same Ultra-Rapid Parallel Brain Dynamics Underpin the Production and Perception of Speech. *Cerebral Cortex Communications,* 2.

FEDORENKO, E., SCOTT, T. L., BRUNNER, P., COON, W. G., PRITCHETT, B., SCHALK, G. & KANWISHER, N. 2016. Neural correlate of the construction of sentence meaning. *Proceedings of the National Academy of Sciences of the United States of America,* 113**,** E6256-E6262.

FELLEMAN, D. J. & VAN ESSEN, D. C. 1991. Distributed Hierarchical Processing in the Primate Cerebral Cortex. *Cerebral Cortex,* 1**,** 1-47.

FITCH, W. T. 2012. Evolutionary Developmental Biology and Human Language Evolution: Constraints on Adaptation. *Evolutionary Biology,* 39**,** 613-637.

FLORIDI, L. & CHIRIATTI, M. 2020. GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines,* 30**,** 681-694.

FONTOLAN, L., MORILLON, B., LIEGEOIS-CHAUVEL, C. & GIRAUD, A. L. 2014. The contribution of frequency-specific activity to hierarchical information processing in the human auditory cortex. *Nature Communications,* 5.

FRISTON, K. J. 2009. The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences,* 13**,** 293-301.

FRISTON, K. J. & KIEBEL, S. 2009. Cortical circuits for perceptual inference. *Neural Networks,* 22**,** 1093-1104.

FRISTON, K. J., KILNER, J. & HARRISON, L. 2006. A free energy principle for the brain. *Journal of Physiology-Paris,* 100**,** 70-87.

FRISTON, K. J., PARR, T. & DE VRIES, B. 2017. The graphical brain: Belief propagation and active inference. *Network Neuroscience,* 1**,** 381-414.

FRISTON, K. J., PARR, T., YUFIK, Y., SAJID, N., PRICE, C. J. & HOLMES, E. 2020. Generative models, linguistic communication and active inference. *Neuroscience and Biobehavioral Reviews,* 118**,** 42-64.

FRISTON, K. J. & PENNY, W. 2011. Post hoc Bayesian model selection. *Neuroimage,* 56**,** 2089-2099.

FRISTON, K. J., SAJID, N., QUIROGA-MARTINEZ, D. R., PARR, T., PRICE, C. J. & HOLMES, E. 2021. Active listening. *Hearing Research,* 399.

FRISTON, K. J., TRUJILLO-BARRETO, N. & DAUNIZEAU, J. 2008. DEM: A variational treatment of dynamic systems. *Neuroimage,* 41**,** 849-885.

GALANTUCCI, B., FOWLER, C. A. & TURVEY, M. T. 2006. The motor theory of speech perception reviewed (vol 13, pg 361, 2006). *Psychonomic Bulletin & Review,* 13**,** 742-742.

GIRAUD, A. L. & ARNAL, L. H. 2018. Hierarchical Predictive Information Is Channeled by Asymmetric Oscillatory Activity. *Neuron,* 100**,** 1022-1024.

GIRAUD, A. L. & POEPPEL, D. 2012. Cortical oscillations and speech processing: emerging computational principles and operations. *Nat Neurosci,* 15**,** 511-7.

GOLDSTEIN, A., ZADA, Z., BUCHNIK, E., SCHAIN, M., PRICE, A., AUBREY, B., NASTASE, S. A., FEDER, A., EMANUEL, D. & COHEN, A. 2021. Thinking ahead: prediction in context as a keystone of language in humans and machines. *bioRxiv***,** 2020.12. 02.403477.

GREENBERG, S., CARVEY, H., HITCHCOCK, L. & CHANG, S. Y. 2003. Temporal properties of spontaneous speech - a syllable-centric perspective. *Journal of Phonetics,* 31**,** 465-485.

GREENFIELD, P. M. 1991. Language, Tools, and Brain - the Ontogeny and Phylogeny of Hierarchically Organized Sequential Behavior. *Behavioral and Brain Sciences,* 14**,** 531-550.

GRIFFITHS, T., STEYVERS, M., BLEI, D. & TENENBAUM, J. 2004. Integrating topics and syntax. *Advances in neural information processing systems,* 17.

GRIFFITHS, T. L., STEYVERS, M. & TENENBAUM, J. B. 2007. Topics in semantic representation. *Psychol Rev,* 114**,** 211-44.

GWILLIAMS, L., KING, J.-R., MARANTZ, A. & POEPPEL, D. 2020. Neural dynamics of phoneme sequencing in real speech jointly encode order and invariant content. *bioRxiv***,** 2020.04.04.025684.

HALE, J. A probabilistic Earley parser as a psycholinguistic model. Second meeting of the north American chapter of the association for computational linguistics, 2001.

HAUSER, M. D., CHOMSKY, N. & FITCH, W. T. 2002. The faculty of language: What is it, who has it, and how did it evolve? *Science,* 298**,** 1569-1579.

HEILBRON, M., ARMENI, K., SCHOFFELEN, J.-M., HAGOORT, P. & DE LANGE, F. P. 2021. A hierarchy of linguistic predictions during natural language comprehension. *bioRxiv***,** 2020.12. 03.410399.

HICKOK, G. & POEPPEL, D. 2007. Opinion - The cortical organization of speech processing. *Nature Reviews Neuroscience,* 8**,** 393-402.

HOVSEPYAN, S., OLASAGASTI, I. & GIRAUD, A.-L. 2022. Rhythmic modulation of prediction errors: a possible role for the beta-range in speech processing. *bioRxiv***,** 2022.03.28.486037.

HOVSEPYAN, S., OLASAGASTI, I. & GIRAUD, A. L. 2020. Combining predictive coding and neural oscillations enables online syllable recognition in natural speech. *Nature Communications,* 11.

KOSKINEN, M., KURIMO, M., GROSS, J., HYVARINEN, A. & HARI, R. 2020. Brain activity reflects the predictability of word sequences in listened continuous speech. *Neuroimage,* 219**,** 116936.

KUPERBERG, G. R. 2007. Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research,* 1146**,** 23-49.

LAKATOS, P., GROSS, J. & THUT, G. 2019. A New Unifying Account of the Roles of Neuronal Entrainment. *Curr Biol,* 29**,** R890-R905.

LECUN, Y. & BENGIO, Y. 1995. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks,* 3361**,** 1995.

LEONARD, M. K., BAUD, M. O., SJERPS, M. J. & CHANG, E. F. 2016. Perceptual restoration of masked speech in human cortex. *Nature Communications,* 7.

LEVINSON, S. E. 1986. Continuously variable duration hidden Markov models for automatic speech recognition. *Computer Speech & Language,* 1**,** 29-45.

LEVY, R. 2008. Expectation-based syntactic comprehension. *Cognition,* 106**,** 1126-1177.

MACGREGOR, L. J., RODD, J. M., GILBERT, R. A., HAUK, O., SOHOGLU, E. & DAVIS, M. H. 2020. The Neural Time Course of Semantic Ambiguity Resolution in Speech Comprehension. *Journal of Cognitive Neuroscience,* 32**,** 403-425.

MAMASHLI, F., KHAN, S., OBLESER, J., FRIEDERICI, A. D. & MAESS, B. 2019. Oscillatory dynamics of cortical functional connections in semantic prediction. *Hum Brain Mapp,* 40**,** 1856-1866.

MARTIN, A. E. 2020. A Compositional Neural Architecture for Language. *J Cogn Neurosci,* 32**,** 1407-1427.

MARTIN, A. E. & DOUMAS, L. A. 2017. A mechanism for the cortical computation of hierarchical linguistic structure. *PLoS Biol,* 15**,** e2000663.

MCCLELLAND, J. L. & ELMAN, J. L. 1986. The Trace Model of Speech-Perception. *Cognitive Psychology,* 18**,** 1-86.

MCRAE, K., FERRETTI, T. R. & AMYOTE, L. 1997. Thematic roles as verb-specific concepts. *Language and Cognitive Processes,* 12**,** 137-176.

MEYER, L., SUN, Y. & MARTIN, A. E. 2020. Synchronous, but not entrained: exogenous and endogenous cortical rhythms of speech and language processing. *Language Cognition and Neuroscience,* 35**,** 1089-1099.

MURPHY, E. 2018. Interfaces (travelling oscillations)+ recursion (delta-theta code)= language. *The Talking Species: Perspectives on the Evolutionary, Neuronal and Cultural Foundations of Language, eds E. Luef and M. Manuela (Graz: Unipress Graz Verlag)***,** 251-269.

NABE, M., SCHWARTZ, J. L. & DIARD, J. 2021. COSMO-Onset: A Neurally-Inspired Computational Model of Spoken Word Recognition, Combining Top-Down Prediction and Bottom-Up Detection of Syllabic Onsets. *Frontiers in Systems Neuroscience,* 15.

NEUROIMAGING, W. T. C. F. 2014. SPM12.

NORRIS, D. 1994. Shortlist - a Connectionist Model of Continuous Speech Recognition. *Cognition,* 52**,** 189-234.

PARK, H., INCE, R. A., SCHYNS, P. G., THUT, G. & GROSS, J. 2015. Frontal top-down signals increase coupling of auditory low-frequency oscillations to continuous speech in human listeners. *Curr Biol,* 25**,** 1649-53.

PARR, T., REES, G. & FRISTON, K. J. 2018. Computational Neuropsychology and Bayesian Inference. *Frontiers in Human Neuroscience,* 12.

PAYNE, J. W., BETTMAN, J. R. & JOHNSON, E. J. 1988. Adaptive Strategy Selection in Decision-Making. *Journal of Experimental Psychology-Learning Memory and Cognition,* 14**,** 534-552.

PEELLE, J. E. 2018. Listening Effort: How the Cognitive Consequences of Acoustic Challenge Are Reflected in Brain and Behavior. *Ear and Hearing,* 39**,** 204-214.

PEFKOU, M., ARNAL, L. H., FONTOLAN, L. & GIRAUD, A. L. 2017. theta-Band and beta-Band Neural Activity Reflects Independent Syllable Tracking and Comprehension of Time-Compressed Speech. *Journal of Neuroscience,* 37**,** 7930-7938.

PULVERMULLER, F. 2018. Neural reuse of action perception circuits for language, concepts and communication. *Progress in Neurobiology,* 160**,** 1-44.

PULVERMULLER, F. & FADIGA, L. 2010. Active perception: sensorimotor circuits as a cortical basis for language. *Nature Reviews Neuroscience,* 11**,** 351-360.

RADFORD, A., WU, J., CHILD, R., LUAN, D., AMODEI, D. & SUTSKEVER, I. 2019. Language models are unsupervised multitask learners. *OpenAI blog,* 1**,** 9.

RAO, R. P. N. & BALLARD, D. H. 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience,* 2**,** 79-87.

RIMMELE, J. M., POEPPEL, D. & GHITZA, O. 2021. Acoustically Driven Cortical δ Oscillations Underpin Prosodic Chunking. *Eneuro,* 8.

RODD, J. M., DAVIS, M. H. & JOHNSRUDE, I. S. 2005. The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity. *Cereb Cortex,* 15**,** 1261-9.

SCHRIMPF, M., BLANK, I. A., TUCKUTE, G., KAUF, C., HOSSEINI, E. A., KANWISHER, N., TENENBAUM, J. B. & FEDORENKO, E. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences of the United States of America,* 118.

SHANNON, C. E. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal,* 27**,** 379-423.

SOHOGLU, E., PEELLE, J. E., CARLYON, R. P. & DAVIS, M. H. 2012. Predictive top-down integration of prior knowledge during speech perception. *J Neurosci,* 32**,** 8443-53.

SWINNEY, D. A. 1979. Lexical Access during Sentence Comprehension - (Re)Consideration of Context Effects. *Journal of Verbal Learning and Verbal Behavior,* 18**,** 645-659.

TANENHAUS, M. K., CARLSON, G. & TRUESWELL, J. C. 1989. The Role of Thematic Structures in Interpretation and Parsing. *Language and Cognitive Processes,* 4**,** Si211-Si234.

TANENHAUS, M. K., SPIVEYKNOWLTON, M. J., EBERHARD, K. M. & SEDIVY, J. C. 1995. Integration of Visual and Linguistic Information in Spoken Language Comprehension. *Science,* 268**,** 1632-1634.

WANG, L., HAGOORT, P. & JENSEN, O. 2018. Gamma Oscillatory Activity Related to Language Prediction. *Journal of Cognitive Neuroscience,* 30**,** 1075-1085.

WARREN, R. M. 1970. Perceptual restoration of missing speech sounds. *Science,* 167**,** 392-3.

WILLEMS, R. M., FRANK, S. L., NIJHOF, A. D., HAGOORT, P. & VAN DEN BOSCH, A. 2016. Prediction During Natural Language Comprehension. *Cereb Cortex,* 26**,** 2506-2516.

YILDIZ, I. B., VON KRIEGSTEIN, K. & KIEBEL, S. J. 2013. From Birdsong to Human Speech Recognition: Bayesian Inference on a Hierarchy of Nonlinear Dynamical Systems. *Plos Computational Biology,* 9.

Appendix

Appendix I. Model parameters for lemma generation

1. Possible states and prior belief (D matrix) for each factor
   a. Context level

   Context = {'poker game', 'tennis game', 'night party', 'racing game'}, Nc = 4

   D{1} = [1.5 1 1 1]

   Type = {'event', 'property'}, NTy = 2

   D{2} = [1 1]

   b. Semantic

   Agent = {'card A', 'serve', 'run', 'card J', 'neckband', 'score', 'buzz'}, Na = 7

   Relation = {'win', 'ruin', 'be'}, Nr = 3

   Patient = {'tennis', 'poker', 'game', 'evening', 'null'}, Np = 5

   Modifier = {'sufficient', 'unexpected', 'not pretty', 'not fair', 'high volume', 'high freq', 'null'}, Nm = 7

   Prior belief for semantic and syntax factors are calculated by multiplying the higher-level priors with probability mapping matrices defined in the next section.

   c. Syntax

   For syntax, we treat each epoch differently as if there is one factor syntax{i} for each epoch.

   Syntax{1-4} = {'attribute', 'subject', 'verb', 'object', 'adjective'}, Nsyn = 5

   d. Lemma

   Lemma = {'one more', …, 'sharp'}, Nw = 20. The matching between lemma and syntax + semantic is defined in Appendix I. Multiple meanings are arranged assuming meaning 1 is the major meaning, meaning 2 & 3 are less likely.

2. Probabilistic mapping for the generative model

   In all mapping matrices, the first dimension represents the outcome factor, the 2nd and further dimensions represent states of the higher level. All matrices are normalized so that the first dimension add to 1.

   a. Context to semantic

   For the mapping from context c and sentence type Ty to a semantic factor s (s={ a', 'r', 'p', 'm'}), the model is defined by a 3-D matrix L{s}(Ns, Nc, NTy). Indices refer to the state of the factor, e.g. L{agent}(1, 1, 2) = p(agent='card A'|context='poker', type='property').

   Context = 'poker game' = 1:

   L{agent}('card A', 1, 1) = L{agent}('card A', 1, 2) = 0.6, L{agent}('card J', 1, 1) = L{agent}('card J', 1, 2) = 0.4

   L{relation}('win', 1, 1) = 1, L{relation}('be', 1, 2) = 1

   L{patient}('poker', 1, 1) = 1, L{patient}('null', 1, 2) = 1

   L{modifier}('null', 1, 1) = 1, L{modifier}('sufficient', 1, 2) = L{modifier}('unexpected', 1, 2) = 0.5

   Context = 'tennis game' = 2:

L{agent}('serve', 2, 1) = L{agent}('serve', 2, 2) = 0.6, L{agent}('run, 2, 1) = L{agent}('run', 2, 2) = 0.4

L{relation}('win', 2, 1) = 1, L{relation}('be', 2, 2) = 1

L{patient}('tennis', 2, 1) = 1, L{patient}('null', 2, 2) = 1

L{modifier}('null', 2, 1) = 1, L{modifier}('sufficient', 2, 2) = L{modifier}('unexpected', 2, 2) = 0.5

Context = 'night party' = 3:

L{agent}('neckband', 3, 1) = L{agent}('neckband', 3, 2) = 0.6, L{agent}('buzz', 3, 1) = L{agent}('buzz', 3, 2) = 0.4

L{relation}('ruin', 3, 1) = 1, L{relation}('be', 3, 2) = 1

L{patient}('evening', 3, 1) = 1, L{patient}('null', 3, 2) = 1

L{modifier}('null', 3, 1) = 1

L{modifier}('not pretty', 3, 2) = L{modifier}('high volume', 3, 2) = L{modifier}('high freq', 3, 2) = 1/3

Context = 'racing game' = 4:

L{agent}('score', 4, 1) = L{agent}('score', 4, 2) = 0.6, L{agent}('buzz', 4, 1) = L{agent}('buzz', 4, 2) = 0.4

L{relation}('ruin', 4, 1) = 1, L{relation}('be', 4, 2) = 1

L{patient}('game', 4, 1) = 1, L{patient}('null', 4, 2) = 1

L{modifier}('null', 4, 1) = 1

L{modifier}('not fair', 4, 2) = L{modifier}('high volume', 4, 2) = L{modifier}('high freq', 4, 2) = 1/3

b. Sentence type to syntax

For syntax, we define the model Z for each epoch τ separately. Each Z{τ} is a Nsyn x NTy matrix, i.e. 5x2

Z{1}(:, 1) = Z{1}(:, 2) = [1 0 0 0 0]. The first epoch is always 'attribute'

Z{2}(:, 1) = Z{2}(:, 2) = [0 1 0 0 0]. The second epoch is always a subject

Z{3}(:, 1) = Z{3}(:, 2) = [0 0 1 0 0]. The third epoch is always a verb

Z{4}(:, 1) = [0 0 0 1 0], Z{4}(:, 2) = [0 0 0 0 1]. The fourth epoch could be either an object (sentence type='event') or an adjective (Ty='property').

Now we can calculate priors for semantic and syntax (matrix multiplication is only demonstrative, not concerning matrix transposing in practice).

Semantic: D{3-6}= L{1-4}*D{1}*D{2}

Syntax: D{7}(:, τ) = Z{τ}*D{2}

c. Semantic to lemma

Because we have a one-to-one correspondence between syntax and semantic, we define the model A with Nsyn independent matrices, each mapping from the corresponding semantic factor to word, therefore Nw x Ns.

Special case for A{1} because it does not need semantic information. Therefore A{1} is a Nw x 1 matrix. A{1}(one more) = A{1}(that) = 0.5

A{2}(index, agent) = 1, where index is calculated by finding the dictionary entry for the corresponding agent. E.g. the 3rd agent state, 'card J', translates to 'joker' in the dictionary, which is at the 5th entry. Therefore A{2}(5, 3) = 1. The rest of semantic-lemma mappings are defined in the same fashion.

When one semantic value corresponds to multiple lemma entries, e.g. the patient 'tennis' (1st patient) can be translated into either 'the tennis' (11th lemma entry), or 'the game' (13th), we give higher probability to the "meaning 1" mapping. A{4}(11, 1) = 0.8, A{4}(13, 1) = 0.2.

Appendix II. The mapping between lemma and semantic in the model's mental lexicon

	lemma	Meaning 1	Meaning 2	Meaning 3
1	'one more'	'extra'	[]	[]
2	'that'	'that'	[]	[]
3	'ace'	'card A'	'serve'	[]
4	'sprint'	'run'	[]	[]
5	'joker'	'card J'	[]	[]
6	'tie'	'neckband'	'score'	[]
7	'noise'	'buzz'	[]	[]
8	'wins'	'win'	[]	[]
9	'ruined'	'ruin'	[]	[]
10	'is'	'be'	[]	[]
11	'the tennis'	'tennis'	[]	[]
12	'the poker'	'poker'	[]	[]
13	'the game'	'game'	'tennis'	'poker'
14	'the evening'	'evening'	[]	[]
15	'enough'	'sufficient'	[]	[]
16	'surprising'	'unexpected'	[]	[]
17	'ugly'	'not pretty'	[]	[]
18	'unfair'	'not fair'	[]	[]
19	'loud'	'high volume'	[]	[]
20	'sharp'	'high freq'	[]	[]

Appendix III. Full expression of free energy and gradient descent algorithm for the top-level model (L_1 and L_2)

The expression of free energy (eq.16 in the main text) can be parameterized with respect to the posterior estimates of L_1 and L_2 factors

$$\begin{aligned} F &= \sum_{\tau,k} syn_k^\tau \ln(syn_k^\tau) + \sum_{\alpha,j} s_j^\alpha \ln s_j^\alpha + \sum_m c_m \ln c_m + \sum_a ST_a \ln ST_a \\ &- \sum_{i,j,k,\tau} W_i^\tau \ln A_{i,j,k}^{(k)} syn_k^\tau s_j^{\alpha(k)} - \sum_{a,k,\tau} \ln Z_{k,a}^\tau syn_k^\tau ST_a \\ &- \sum_{j,a,m,\alpha} \ln L_{j,m,a}^{(\alpha)} s_j^\alpha c_m ST_a - \sum_a \ln H_a ST_a - \sum_m \ln D_m c_m \end{aligned}$$

We can then derive partial derivatives of F with respect to Q:

$$\begin{aligned} \frac{\partial F}{\partial syn_k^\tau} &= \ln syn_k^\tau - \sum_{i,j} W_i^\tau s_j^{\alpha(k)} \ln A_{i,j,k}^{(k)} - \sum_a ST_a \ln Z_{k,a}^{(\tau)} \\ &\quad \tau = 1, \cdots, n \\ \frac{\partial F}{\partial s_j^\alpha} &= \ln s_j^\alpha - \sum_i (W_i^2 syn_k^2 + \cdots + W_i^n syn_k^n) \ln A_{i,j,k}^{k(\alpha)} \\ &\quad - \sum_{m,a} c_m ST_a \ln L_{j,m,a}^\alpha \\ &\quad \alpha = \{A, R, P, M\} \\ \frac{\partial F}{\partial c_m} &= \ln c_m - \sum_{j,a} s_j^{(1)} ST_a \ln L_{j,m,a}^{(1)} - \cdots \\ &\quad - \sum_{j,a} s_j^{(n_s)} ST_a \ln L_{j,m,a}^{(n_s)} - \ln D_m \\ \frac{\partial F}{\partial ST_a} &= \ln ST_a - \sum_k syn_k^{(1)} \ln Z_{k,a}^{(1)} - \cdots - \sum_k syn_k^{(n)} \ln Z_{k,a}^{(n)} \\ &\quad - \sum_{j,m,\alpha} \ln L_{j,m,a}^{(\alpha)} s_j^\alpha c_m - \ln H_a, \quad \alpha = \{A, R, P, M\} \end{aligned}$$

To solve the above equations, we follow Friston et al. (2017) and define an auxiliary variable *v* for the estimation of each factor *x*. Let $x \equiv \sigma(v)$, where σ() denotes the softmax function. We can then solve *v* and *x* using gradient descent:

$$\dot{v}_k = -\frac{\partial F}{\partial x_k}$$

With this choice for $\dot{v}_k$, the update equations are

$$\begin{aligned} v_k(t + \Delta t) &\approx v_k(t) - \Delta t \frac{\partial F}{\partial x_k} \\ \vec{x}(t + \Delta t) &\equiv \vec{\sigma}(\vec{v}(t + \Delta t)) \approx \vec{\sigma}\left(v_k(t) - \Delta t \frac{\partial F}{\partial x_k}\right) \end{aligned}$$