Neuro-computational models of language processing

John T Hale, Luca Campanelli, Jixing Li, Shohini Bhattasali, Christophe Pallier, Jonathan R Brennan

Neuro-computational models of language processing

John T. Hale,[1] Luca Campanelli,[1,2] Jixing Li,[3] Shohini Bhattasali,[4] Christophe Pallier[5] and Jonathan R. Brennan[6]

[1]Department of Linguistics, University of Georgia, Athens, GA 30602, USA; email: jthale@uga.edu
[2]Haskins Laboratories, New Haven, CT 06511, USA
[3]Neuroscience of Language Lab, NYU Abu Dhabi, United Arab Emirates
[4]University of Maryland, College Park, MD 20742, USA
[5]Cognitive Neuroimaging Unit, INSERM U992, 91191 Gif-sur-Yvette Cedex, France
[6]Department of Linguistics, University of Michigan, Ann Arbor, MI 48109, USA

Xxxx. Xxx. Xxx. Xxx. YYYY. AA:1–23

https://doi.org/10.1146/((please add article doi))

Keywords

neurolinguistics, brain, computational model, deep learning, parsing, lexicon

Abstract

Efforts to understand the brain bases of language face the mapping problem: at what level do linguistic computations and representations connect to human neurobiology? We review one approach to this problem that relies on rigorously defined computational models to specify the links between linguistic features and neural signals. Such tools can be used to estimate linguistic predictions, model linguistic features, or specify a sequence of processing steps that may be quantitatively fit to neural signals collected while participants use language. Progress has been helped by advances in machine learning, attention to linguistically interpretable models, and openly shared datasets that allow researchers to compare and contrast a variety of models. We describe one such dataset in detail in the supplementary materials.

Contents

1. Linguistics as abstract neurobiology

"Verbing weirds language" says the boy Calvin to his friend Hobbes in a 1993 newspaper comic strip written by Bill Watterson.[1] The thing is, Hobbes (a stuffed tiger made animate by Calvin's imagination) has no problem understanding this sentence, and nor of course does the reader. Language users regularly deal with the abstractions of human language, whether turning adjectives into verbs, as in the comic strip, formulating a question from a declarative statement, or recognizing the dual meanings of the statement "Yesterday I saw an elephant in my pajamas." Linguists make sense of these patterns by recognizing that the human capacity for language rests on a complex set of abstractions: from phonological features to syntactic categories, Linguistics is abstract. As with other cognitive faculties, these linguistic abstractions are instantiated in the concrete neurobiology of language users.

The biological foundation of linguistic knowledge has been a pillar of generative linguistics since the 1960's (Lenneberg 1967). Yet, efforts to map between the abstract constructs of linguistics and concrete properties of the human brain face significant challenges. Embick & Poeppel (2015) discuss two specific barriers to grounding linguistics in neurobiology. The first concerns ontology, or the entities that are discussed in theories of linguistics versus theories of neuroscience: How does a noun relate to a neuron, or how do patterns of electrical discharge relate to syntactic dependencies? It is not plausible to simply draw a line from a primitive in one domain to a primitive in another. The second barrier concerns granularity: what is the "right" kind of primitive for connecting linguistic abstractions with neurobiological properties? Are there specific neuronal columns devoted to computing noun phrases, or do phrasal categories emerge from interactions between disparate cell assemblies? Is the noun phrase an appropriate unit to connect with neuronal function, or ought linkings be made to specific constructions, or perhaps to more general notions such as Merge? The field has not reached a consensus about how to tackle these questions, collectively called the *Mapping Problem* between linguistics and neurobiology (Poeppel 2012).

The Mapping Problem: The challenge of linking linguistic and neural primitives in an explanatory way.

This paper reviews one approach to answering the Mapping Problem, one that relies on computational models of language processing. These models explicitly specify how properties of human language relate to real time processing, and also how such processes might affect observable neural signals. As such, this approach relates both to concrete properties of the brain and to live issues in the study of human language.

[1] *Calvin and Hobbes* - January 25, 1993.

Box 1: Bluffer's Guide to Neurolinguistic Methods

fMRI Functional Magnetic Resonance Imaging measures the flow of oxygenated blood inside the brain; more oxygenated blood flows in to serve the metabolic needs of neurons that are active, but with a delay of 4-8 seconds. Recordings from fMRI are thus blood-oxygen level dependent, and so commonly referred to as the BOLD signal. This tool offers high spatial but low temporal resolution. See Brennan (2020) or chapter 2 of Kemmerer (2014).

EEG Electroencephalography measures electrical potentials on the scalp that are generated by tens of thousands of neurons from the cortex. These "brainwaves" are sensitive with high temporal resolution to the activity of synchronized neurons that share a common orientation, but the electrical signal is spatially imprecise (see Luck 2014 or page 56ff of Kemmerer 2014 chapter 2).

ERP Event-related potentials are measured by averaging together EEG signals that have been aligned to an event such as stimulus presentation. Well-studied ERP components include the N400, a negative-going central-posterior voltage potential that increases in amplitude for semantically unexpected items, and the P600, a positive-going posterior potential that is related to violated syntactic expectations. See Swaab et al. (2012) for a review.

MEG Magnetoencephalography measures the magnetic fluctuations associated with electrical activity. Like EEG, this signal tracks neural activity as it happens, millisecond by millisecond, but it also offers higher spatial resolution as magnetic fields can be used to triangulate the electrical "source" of activity in the brain. See Salmelin & Baillet (2009) for an overview to this technique, or Hansen et al. (2010) for a textbook introduction.

MVPA Multivariate pattern analysis refers to a diverse set of methods that analyze neural responses as patterns of activity that reflect the varying brain states. MVPA-based decoding analysis makes use of machine learning classifiers, whose inputs are vectors of observed neural signal levels, to guess whether a particular brain state falls into one of a finite number of discrete categories. These categories may label linguistic content in the stimuli that were presented to human participants. If the classifier achieves above-chance performance, then researchers infer that information corresponding to that category label was present within the brain, at the place and time the neural signals were taken (see Pereira et al. 2009).

2. The scope of this review

As suggested above, the key element is computation. It is helpful to distinguish between the scientific use of computers in general and the computational theory of mind as a foundation for cognitive science. While computers may serve to deduce the consequences of linguistic theory or analyze neural signals, these sorts of computations are not the focus of this review. Rather, we focus on computations and representations that play an explanatory role in answering the question "how does an individual use language?" This

garden-variety computationalism has been a meeting-point for linguists and brain scientists over the years (for a philosophical overview see Rescorla 2020).

The present article specifically considers neuro-computational models that match time-series brain data in order to shed light on questions of interest to linguists. It updates Stowe et al. (2005) with computer models of language processing in particular regions of the brain. We consider neurolinguistic methods involving naturalistic text stimuli which have already been reviewed in Brennan (2016), and we touch on classifier-based decoding of language in the brain, which is introduced in Murphy et al. (2018).

The current review sets aside several closely related areas. One is the research tradition rooted aphasiology, which is often computational in exactly the sense evoked above (e.g. Caplan 1992). Another closely-related area is computational psycholinguistics; for a general review of that field we refer the reader to Hale (2017). We also set aside research using signals that reflect neural oscillations (e.g. Bastiaansen & Hagoort 2006; Meyer et al. 2020), while recognizing the promise of that direction of research when combined with explicit computational models of the kind considered in this review (Martin & Doumas 2017).[2]

As Poeppel (2012, already cited above) emphasizes, scientific progress hinges upon our ability to explicitly link theorized linguistic concepts, such as phrase structure, and observable neurobiological phenomena, such as blood flow in the brain. These sorts of connections between theory and data inevitably reflect additional assumptions about cognitive processing. Using the methodologies in box 1, the next sections sketch several specific approaches to building such links. Current work using these methods is assessed with respect to its advantages and disadvantages for linguistics.

3. Guessing the next word or symbol

A prominent perspective on the brain construes it to be an "inference engine" that optimizes predictive accuracy (Friston 2010). Neural signals, such as those described in box 1, certainly show exquisite sensitivity to deviations from an expected outcome. This sensitivity has been leveraged to probe the sorts of representations in use during language processing. In this type of work, computational models serve to define probability distributions over upcoming linguistic input.

The surprisal linking hypothesis A probability distribution over upcoming linguistic expressions quantifies how predictable those expressions are. This degree of predictability can be formalized as the self-information or *surprisal* of an event, such as a particular successor word following a given left-context string (see Hale 2016, for a review). One way to estimate such a probability distribution is by counting. For instance, the probability of the word "snow" in the phrase "look at the falling snow" can be estimated by counting in a large corpus of text how often the words "the falling" are followed by "snow." Counting word sequences by threes in this way is a trigram Markov Model (see box 2). This approach was taken by Willems et al. (2015) to work out numerical surprisal values that quantify the unexpectedness of each successive word. Using fMRI they find increased activation for higher surprisal in the superior temporal lobe of the left and right hemisphere.

Surprisal: Deviation between an expected and actual linguistic input; equivalent to the bits of information conveyed.

This research follows Brennan et al. (2012) and Yarkoni et al. (2008) in relying on neu-

[2] See also the commentaries collected in *Language, Cognition & Neuroscience*, 2020, Volume 35 Issue 9.

Box 2: Language models

A language model assigns probabilities to sentences and their substrings.

ngram An ngram language model uses the previous $n-1$ symbols as a conditioning context for the probability of the n^{th} symbol, a Markov assumption. The symbols may come from any discrete set including letters, phones, words, or parts of speech. The probabilities are estimated from corpora and are typically "smoothed" to handle unseen events (see Eisenstein 2019, §6.1, 6.2). The case $n = 3$ is called a trigram model.

CFG Context-Free Grammar, also called Phrase-Structure Grammar, is a formal model of language based on symbolic rewriting rules. These symbols and rules induce equivalence classes over linguistic expressions, for instance noun phrases with or without an optional adjective. Such grammars served as the "base component" in transformational grammar (see Partee et al. 1993, chapter 16). Adding weights to CFG rules yields probabilistic context-free grammars (PCFG) whose derivations can be treated as a stochastic branching process (see Manning & Schütze 2000, chapter 11), from which a language model can be derived (Jelinek & Lafferty 1991).

RNNs In recurrent networks, the same block of neural net architecture is "unrolled" so that it applies over and over again to each position in a string. This allows it to adaptively learn from examples using a variant of the backpropagation algorithm (Werbos 1990). Recurrent nets based on gated connections such as the Long Short-Term Memory of Hochreiter & Schmidhuber (1997) fare better than their ungated counterparts when it comes to long-distance dependencies (see Goldberg 2017, chapter 15).

Transformers Transformers are neural net architectures for sequence processing that use an "attention" mechanism based on multiplicative gating (Vaswani et al. 2017). This allows distant symbols to influence the choice of fill-in-the-blank answers. They are typically trained using a masked language modelling objective which allows for influences from either left or right context. Transformers are built to maximize parallelism on modern computers, which entails an upper limit on the length of strings that they can handle (i.e. 512 tokens in BERT).

ral signals derived from "everyday" language processing (for discussion, see Brennan 2016). These sorts of studies elicit ecologically-natural language comprehension by presenting human participants with narrative language much as one might read a story or listen to an audiobook. At issue in subsequent statistical analysis is the degree to which the observed neural signals fit or fail to fit theoretical predictions about the processing difficulty of that text. A key requirement on predictors is broad coverage: the more stimulus words there are that receive a numerical difficulty prediction, the better the chance of being able to estimate a relationship between the predictor and the measured neural signal (e.g. BOLD see box 1). Language models such as the trigram mentioned above in connection with Willems et al.'s 2015 study are "broad coverage" in exactly this sense. But there are other language models that retain the trigram's breadth of coverage while at the same time sharpening the prob-

ability distribution over successor words[3]; some key examples are summarized in box 2. Modern artificial neural network language models fall into this category (see Jurafsky & Martin 2021, chapter 7 or Goldberg 2017, chapter 9). Building on Elman (1990), Frank et al. (2015) shows that neural network language models can account for a well-known expectancy effect in human electrophysiology, the N400 (see box 1).

Surprisal as a degree of change to internal representations Ettinger et al. (2016) and Rabovsky et al. (2018) also model the N400 ERP component. In this line of work, the linking hypothesis is defined in terms of a componentwise changes to a neural network's internal representation. If these components were to receive an interpretation in terms of next-word distributions, then the proposal would reduce to (word) surprisal, as introduced above. But importantly for Rabovsky and colleagues, the representations inside their model are not just about word tokens. Rather they pair up a form of Case Grammar (Fillmore 1968) with a decompositional semantics in the style of Katz & Fodor (1963). Words and sentences are both represented using microfeatures that are *conceptual* in nature, following McClelland & Kawamoto (1986) and Rohde (2002). This kind of modeling fleshes out, in computational detail, the established view of the N400 as having to do with dashed meaning expectations (Kutas & Federmeier 2000).

Brouwer et al. (2021) extend this general approach in two ways. First, they model the P600 effect as well as the N400. Secondly, they do so using an internal representation that relates to models of propositional logic. By providing a 'microworld' interpretation for vector-valued internal states of a neural network, Brouwer et al. aim to capture the effect of real-world knowledge on human sentence processing.

The degree to which intermediate states of the human language comprehender should be viewed as "model-like" (as opposed to proof-like) remains open (see e.g. Hemforth & Konieczny 2006). So too is the question of whether or not neural sequence models should be thought of as possessing knowledge of grammar (see e.g. Steedman 1999; Linzen et al. 2016).

Structures & sequences Neural network sequence models of the type mentioned above share with ngram models a conceptualization of language as a time series of words, lacking any hierarchical structure. Chomsky's 1956 critique of this beads-on-a-string outlook is all too familiar to linguists. Cognizant of this critique, other researchers have used hierarchical syntactic analyses as an intermediate representation for modelling time series brain data.

For instance Brennan & Hale (2019) demonstrate that a phrase-structured model of part-of-speech tag sequences yields a better model of EEG data than does an ngram model or simple recurrent network in the style of Frank et al. (2015, already cited above). This project used the surprisal linking hypothesis. It uncovered a different kind of electrophysiological response, one that manifests earlier and on electrodes that are more anterior than the usual N400. Others have similarly confirmed the explanatory role of hierarchical syntax in other methodologies.

To take just one fMRI example, Shain et al. (2020) use the surprisal linking hypothesis to work out processing difficulty predictions from a different sort of probabilistic phrase-structure grammar (see also Henderson et al. 2016). They also derive surprisals from a

[3]These next-word probability distributions depend strongly on the genre-specific characteristics of their training data (see e.g. Hale et al. 2019).

5-gram Markov Model, one whose parameters can only be estimated with exceptionally large corpora. This point highlights a challenge in extending heavily data-reliant methods to broader cross-linguistic study, especially of under-resourced languages. Shain et al. find that these two sorts of surprisal, one based on hierarchical phrase-structure, and another based on word sequences, explain independent variability in the BOLD signal in a range of areas through the "language network" of the frontal and temporal lobes in the left hemisphere. That is, the data are consistent with processing mechanisms that are sensitive to hierarchical dependencies along-side local word-to-word transitions.

These studies reason backwards from the goodness of fit between fMRI data and hierarchically-structured language models to the scientific claim that hierarchical structure really is part of human language comprehension. This reasoning relies upon a comparison with a sequence-oriented, hierarchy-free language models. Hale et al. (2018) and Brennan et al. (2020) look at a subtler comparison, one that pits bona fide phrase structure composition against a "composition-free" kind of parsing. Bona fide phrase structure composition is formalized using the Recurrent Neural Network Grammars of Dyer et al. (2016, RNNG). The baseline is a language model whose outputs are not syntactic trees but merely bracketed strings. Neural networks using this "non-compositional" alternative representation (inspired by Vinyals et al. 2015 and Choe & Charniak 2016, see the discussion of finite state transducers in Langendoen 2008) are not compelled to form truly hierarchical internal representations. Statistical comparison of the surprisal values from the two models as regards human EEG highlights a time window about 200 to 300 milliseconds after a word is encountered when explicit hierarchical composition appears to be happening.

Multi-level prediction Varying the symbols over which a language model is defined can also lead to insight. For instance Lopopolo et al. (2017) use a quantity that is closely related to surprisal, the perplexity of the next symbol at three different levels of abstraction. The first level is phonological, where perplexity is defined in terms of the sequence of speech sounds. At the second level of abstraction, there is the familiar sequence of words. And at a third, the sequence contains part of speech tags. Using a trigram model at each of these levels, Lopopolo and colleagues identify three distinct collections of brain areas. Many of these are clustered in well-known perisylvian language regions but using set-theoretic operations these authors were able to identify, for instance, the Angular Gyrus as a region that simultaneously processes grammatical and phonological information.

In their EEG and MEG study, Heilbron et al. (2021) decompose word predictions from a transformer neural network (GPT-2) into different levels of linguistic analysis. First, like Lopopolo et al., they find dissociable patterns with morphosyntactic and phone-level predictions related to earlier responses (100-400 ms) in, mostly, temporal areas, while lexical-semantic predictions drive later responses ($>$ 400 ms) across a wider set of cortical areas. Second, neural responses at the level of speech sounds are modulated in a top-down way by predictions regarding word meaning. This supports theories in which linguistic predictions not only take place at multiple levels of abstraction, but also modulate each other.

The latter point is consistent with previous findings from the literature. To mention just one, Dikker & Pylkkänen (2013) presented participants with picture-word sequence pairs in which the two elements of the pair could either match (high predictability; e.g., a picture of an apple followed by the word 'apple') or not match (low predictability; e.g., a picture of a banana followed by the word 'apple'). Participants were asked to indicate whether the word described the picture. Interestingly, Dikker & Pylkkänen find that noun

predictability modulated MEG brain responses not only in fronto-temporal areas more directly associated with lexical access but also in lower level sensory areas like the visual cortex. All of these results accord with an "interactive" model of sentence processing, in which different linguistic levels modulate or co-constrain each other (Marslen-Wilson 1975).

4. Linguistic features within and across languages

The Decoding paradigm An alternative response to Poeppel's 2012 Mapping Problem views neural activity not as a indirect indication of work, but a direct characterization of a concept, linguistic or otherwise. This approach, pioneered by Tom Mitchell and Marcel Just is known as MVPA (see box 1). It treats BOLD levels as vector representations of concepts that are extensionally presented in the course of an fMRI experiment. Murphy et al. (2018) review this tradition, which begins with Mitchell et al. (2008). The central idea is to use a kind of learning program called a classifier to discover regularities within the brain data that align with distinctions between the experimental stimuli. These distinctions (say between the lexicalized concepts CELERY and DOG or between Double Object vs Prepositional Dative phrase structures) simply become labels or features for classifiers of brain data. A key innovation in this approach is the use of held-out data; typically the observations are divided up into a training set and a testing set in order to assess the generalization performance of trained classifier. The results can be averaged across alternative divisions to make best use of the available observations while carefully avoiding circular reasoning ("cross-validation", a standard practice in machine learning). The spatial and/or temporal locality of the data can be leveraged to look for regions that are diagnostic for particular types of labels ("searchlight" see Kriegeskorte et al. 2006).

Classifier: Statistical tool that learns to classify, or "decode", a stimulus feature based on brain signals.

However there are limitations, too. Decoding depends on a finite number of trials, and there are correspondingly a finite number of distinctions that can be drawn between all the labelled cases. It's always possible that a high-performing classifier has latched on to a pattern in the brain data that is accidentally, rather than lawfully, related to the intended meaning of the experimenter's label. This limitation is not specific to the decoding paradigm but is rather an instance of the classic problem of Construct Validity in psychology (Cronbach & Meehl 1955; Smith 2005). We take up this concern as it applies specifically to neuro-computational modeling in the conclusion.

Lexical semantics In the seminal 2008 study mentioned above, those concepts were noun meanings, operationalized as "word embeddings", or vectors of co-occurrence counts with a handful of special verbs. This "distributional" approach generalizes the semantic markerese of Katz & Fodor (1963) already mentioned above (see Baroni et al. 2014, on distributional models of meaning). A follow-up study argued that these representations are conceptual, and not lexical, by showing that they generalize across languages such as English and Portuguese (Buchweitz et al. 2012).

Embedding: Vector representation of a word or other linguistic unit.

Huth et al. (2016) apply the same method to natural speech samples (in their case, excerpts from a popular podcast). Forming a distributional vector space based on word-co-occurrence, they observe that distinct lexical-semantic dimensions such as "social", "visual", or "numeric" map to distributed patterns of activity across diverse areas of the cortex. Opening the black box a little bit, they compare their word embeddings favorably against other embeddings that are derived from a trained neural network (word2vec; see Jurafsky &

Martin 2021, chapter 6 or Goldberg 2017, §10.4 for an information-theoretic interpretation).

One approach to scaling these efforts to larger expressions is to simply average the single word vectors (Pereira et al. 2018; see also Li et al. 2016). However this "bag-of-words" approach discards structural information, such as semantic roles. Tensor Product representations (Smolensky & Legendre 2006) and DisCoCat (Coecke et al. 2010) are mathematical theories that offer ways of composing word embeddings into larger structures (see also Baroni et al. 2014).[4] Another less-principled option is to use numerical vector representations that arise via training on tasks such as classifying entailment relations, or guessing a blanked-out word in a sentence. The assumption is that these vectors will encode whatever structural information is truly useful in these auxiliary tasks (e.g. Anderson et al. 2021). However, the details of particular training sets sometimes mean that there are superficial tricks a network can use to succeed at these sorts of auxiliary tasks without actually learning much in the way of sentence-structural information.[5] To avoid this uncertain inference from success on an auxiliary task to adequate linguistic representation with some set of vectors, structural features may also be encoded directly, as discussed in the following subsection.

Syntactic structure and thematic roles Sticking with the one-word, one-vector methodology described above, Wehbe et al. (2014a) extend the Mitchell/Just decoding paradigm to a broader range of linguistic features and structures that characterize naturalistic texts. Wehbe and colleagues assigned features to each word of chapter 9 of *Harry Potter and the Sorcerer's Stone*. These include referential features, i.e. about the identity of particular story characters as well as tags for the part of speech and choice of dependency-syntactic relation borne by a particular word. Human participants read the story one word at a time on a screen. This word at a time character of the presentation facilitates an analysis that considers each brain volume separately, relieving the analyst from any need to assume a particular hemodynamic response function (as in Brennan et al. 2012, 2016; Li & Hale 2019 but cf. Shain et al. 2020). Classifiers are evaluated using a scheme which forces the system to differentiate between a passage's actual brain-data and a nonmatching 'foil' passage's brain data. As before this is done using held-out data in a cross-validated manner to assess whether any generalization beyond the training set would occur.

Wehbe et al.'s innovative study used simple textual features that are not controversial within linguistics. Its findings are consistent with the view that Broca's area subserves the recognition of grammatical dependencies such as Subject, Object, or Modifier (see e.g. Friederici 2017). Zhang et al. (2021) push this MVPA approach in a referential direction, decoding the identity of story characters from naturalistic stimuli, even in "pro-drop" cases where there is no overt mention but instead they would be inferred from context by native speakers.

The promise of the approach lies in its synergy with artificial neural networks. By using

[4]Tensor Product representations have been criticized for matching human similarity judgments less-well than other competing systems (Mitchell & Lapata 2010; Martin & Doumas 2019) and for being too expensive in terms of the required number of artificial neurons. See Smolensky & Legendre (2006, chapter 7) annd Smolensky (2015) for replies.

[5]Caucheteux et al. (2021) introduce an intriguing alternative approach: average neural net activation vectors that arise when applied to ten different sentences that all share the same syntactic structure. In other words, cluster the states of deep neural nets by syntactic equivalence classes. These authors confirm via subsequent probing tasks that syntactic properties are indeed inferrable from these averaged representations.

regularization techniques such as ridge regression, it is possible to learn a linear relationship between (a) internal representations within neural networks and (b) observed brain data such as BOLD signals. This has been done with recurrent neural network sequence models in MEG (Wehbe et al. 2014b) and fMRI (Qian et al. 2016, with a somewhat simpler regression model). Later work extends this to the well-known Jabberwocky manipulation, which neuroscientists use to disentangle syntax from lexical semantics (see e.g. Stowe et al. 1998; Pallier et al. 2011; Matchin et al. 2016; Hashemzadeh et al. 2020) and to Transformers, a high-performing neural net building block for natural language processing (Toneva & Wehbe 2019).

Results to date seem to confirm Stowe et al.'s 2005 suggestion that language processing involves both hemispheres and that it is organized in a processing cascade that proceeds from individual words to larger expressions (cf. Lerner et al. 2011). In virtue of their broad coverage Wehbe et al. demonstrate how that earlier work generalizes to more naturalistic language. They also open the door to explicit comparisons, analogous to Huth et al.'s 2016 comparison between their own word embeddings and embeddings from word2vec (see page 8). Anderson et al. (2021) demonstrate how this work might procede for more abstract structures, by comparing fits between multiple lexical and sentence-based vector representations.

The neural network models mentioned above form their own internal representations, in effect performing a corpus analysis on their training data via the backpropagation algorithm (Rumelhart et al. 1986a). As an alternative to this kind of discovery procedure one can instead start from linguistic features that are given in advance by an experimenter. Allen et al. (2012) did this by decoding neuroimages from human participants who had either read a verb in a prepositional dative structure ("gave a book to John") or a double-object structure ("gave John a book"). Another study by Frankland & Greene (2015) identified regions in the posterior temporal lobe that support above-chance decoding for the semantic roles Agent and Patient. This involved decoding which noun, out of four possible nouns, in fact filled a particular semantic role. Reddy & Wehbe (2021) encode phrase structure trees as numerical vectors in an approximate way, by sampling random walks that traverse those trees node by node. These encodings account for human fMRI observations in well-known language regions. Particularly when the trees include unexpanded top-down predictions about upcoming words, the resulting vectors capture aspects of the fMRI signal above and beyond univariate parser step counts (see section 5 below). These studies all share with Wehbe et al. (2014b) the use of linguistic features that are not controversial. Yet, the underlying methods can be further extended to test features that are more controversial, such as the thematic structure of experiencer predicates (Pesetsky 1995).

Typological comparison As persuasively argued by Ina Bornkessel-Schlesewsky and Matthias Schlesewsky (2013; 2016), typological differences between languages may correspond to different brain networks or different modes of operation of the same brain network, when it comes to language comprehension. Consistent with a universalist view, Honey et al. (2012) find that roughly the same brain regions show BOLD activity when native speakers of English or Russian perform a naturalistic task of comprehending stories in their native language. This converges with other work based on non-naturalistic methods such as Crinion et al. (2006) and Correia et al. (2013). Whereas Honey and colleagues do not identify the specific typological feature driving their result, Dunagan et al. (2021) compares French and Chinese along a single feature: the way a language expresses singular versus plural

number. In French, this marking is overt and obligatory whereas in Chinese it is optional. The results again support the postulation of one common set of language-relevant brain regions that subserve comprehension of number marking, regardless of how it is signaled within a language.

The universalist view supported by these data remains underspecified as to how shared processing resources are recruited by different language varieties. Neuro-computational models, especially those with interpretable components as discussed in the next section, can help unravel how brain mechanisms may be allocated differently across typologically diverse languages. How to specify these cross-linguistic models remains an urgent, open problem.

5. Opening the black box to understand the mechanism

The linguistic features at issue in the previous section are outputs of syntactic and semantic computations that have been performed by candidate models of human language processing. This approach, as well as the prediction-based measures in section 3, are all *extensional* in nature. That is to say, we look at a model's outputs or a probability distribution over such outputs, and compare that theoretical prediction to human brain data in a way that is divorced from the internal process by which the model derived that prediction. This practice has become standard in what Marco Baroni calls "linguistically-oriented deep net analysis" (see Baroni 2021; Linzen & Baroni 2021). Such extensional approaches make it possible to evaluate very large artificial neural nets, ones whose mode of operation defies succinct description. Of course, this cannot be the last word – we cannot simply stop having replaced one inscrutable black box (the brain) with another (deep neural nets). A better response to the mapping problem addresses not only outputs but also internal processing steps that make sense with respect to existing cognitive and linguistic theories.

Derivation steps dissociate from predictions One family of approaches ties directly to processing steps by tapping into parsing models discussed in section 3; these draw on properties of the parser itself, rather than the output representation. For example, the structure-based RNNG parser used by Hale et al. (2018) and Brennan et al. (2020), which was already discussed in terms of surprisal on page 7, may also be queried more directly in terms of the algorithm it implements. Working incrementally from one word to the next, that parsing algorithm includes a loop that iterates over possible syntactic analyses that are consistent with some partial linguistic input. Counting the number of iterations explored between words thus reflects processing effort in a direct way by summing the number of parser actions taken between words. This particular approach builds on a venerable tradition that ties cognitive load to properties of syntactic structure (Kaplan 1972; Frazier 1985). Indeed Brennan et al. (2020) find that such a measure fits with BOLD-signal recorded from the left temporal lobe and inferior frontal gyrus. That improvement in fit is seen only when the parser explicitly composes words into phrases (see page 7) and, moreover, the effect of parser actions is statistically independent of effects that reflect predictability. In other words, while an approach such as this demands a joint commitment to representational and algorithmic elements of a processing model, the relationship of these components to neural signals can be distinguished in a way that guides further, more focused, efforts.

Comparing process models Nelson et al. (2017) takes on the challenge directly of comparing different process models while holding constant representational content (see also Brennan & Pylkkänen 2017). That work made use of exceptionally high-resolution data collected using inter-operative recordings from neurosurgery patients. These recordings come from electrodes placed directly onto the cortex to aid in the surgical treatment of epilepsy; this placement leads to data that can distinguish neural signals with a high degree of both spatial and temporal precision. Nelson et al. presented such patients with sentences that spanned a range of syntactic constructions. The sentences were generated by a context-free phrase-structure grammar (see box 2). A set of process models were constructed that differed in how eagerly, or predictively, they traversed these phrase-structures. These include the least eager "bottom-up" strategy which draws tree nodes only when all associated leaf nodes have been incrementally encountered, in contrast to the "left-corner" and "top-down"; strategies which each posit syntactic nodes more predictively (cf. Abney & Johnson 1991).

Correlations between the number of parser actions under each of these strategies with the electrophysiological signals reveal a cline such that the left-corner and bottom-up strategies show superior fits to left temporal data than the most eager top-down strategy. These differences are not equally distributed across language-related temporal and frontal regions, indicating that the brain bases of parsing must be understood in a more nuanced way as implementing a variable parsing strategy that is more eager in one region than another. Or there may be construction-specific variation such that certain linguistic expressions are understood in a more eager manner than others.

Process models for realistic grammars In mathematical linguistics, the degree of "expressivity" required for an adequate syntactic theory has been largely resolved: human grammar must be more than context-free, but not a lot more. Stabler (2013) lays out this hidden consensus in an accessible way, building on seminal work by Joshi (1985) and Joshi et al. (1991).[6] Kallmeyer (2010) offers a textbook introduction to parsing algorithms for these formalisms. Does a grammar whose level of expressivity is well-calibrated to human grammar *in general* yield processing-difficulty predictions that fit a *particular person's* brain? Brennan et al. (2016) and Li & Hale (2019) offer an affirmative answer. They find that node-counts on X-bar structures derived by Minimalist Grammars predict unique variance in BOLD signal in the posterior temporal lobe, variance that is not captured by ngrams or naïve phrase structure in the style of the Penn Treebank (see Marcus et al. 1993).[7] Stanojević et al. (2021) go farther, operating directly on the derivations assigned by a near-context free formalism, Combinatory Categorial Grammar (CCG). They find that CCG improves BOLD modeling in six language-relevant brain regions. Adding a special parser action that facilitates late-attachment of modifier phrases improves the fit in both anterior temporal lobe and inferior frontal gyrus ("Broca's Area"). It would be premature to conclude on the basis of just these results that CCG or Minimalism is uniquely the right theory of human grammar. However, they do suggest that these competence theories capture distinctions that play some role in human language processing.

[6] This consensus spans Minimalism (Stabler 2001), Combinatory Categorial Grammar (§8.3 Steedman 2000) and Lexical-Functional Grammar (Wedekind & Kaplan 2020) to name just three frameworks.

[7] This region where Minimalist Grammars make a distinguishable contribution is consistent with Wernicke's area.

Storage versus parsing A perennial question in linguistics concerns the distinction between atomic expressions that are simply remembered whole and structured expressions that are composed from smaller pieces (for a study with non-naturalistic stimuli see Lyu et al. 2019 or Pylkkänen 2019 for a review of the composition operation). Bever's (1970) classic idea is that some expressions are understood as a matter of memory association; the link between a morpheme sequence and a functional label such as Actor-Action-Object is nothing less and nothing more than a habit. This idea has enduring appeal but there remains the issue of exactly where to draw the line between habits and rules (Townsend & Bever 2001; O'Donnell 2015; Blache 2018). Bhattasali et al. (2019) approach this question from standpoint of quantitative linguistics using naturalistic stimuli and employing statistical association metrics (see e.g. Evert 2009) to quantify the "stickiness" of word sequences that may correspond to memorized expressions. Entering these association metrics as predictors into a linear regression as in section 3 highlights the precuneus, a brain region that has been implicated in episodic memory. This result can be taken as supporting a variation on Bever's proposal, namely that human comprehension is interleaved such that sometimes memorized chunks are directly recalled in a way that depends on brain networks involving the precuneus and other times composed in a way that involves the temporal lobe (cf. Ullman 2004; Hagoort 2016).

Reference and coreference Of course quite apart from their internal structure, expressions in a language have reference – they refer to people, places and things. Li et al. (2020) studied this referential aspect of language using complexity metrics to rate the processing difficulty of particular coreference links, within a single literary narrative. As in section 3 this narrative served as an auditory stimulus for listeners who were undergoing fMRI. Across both English and Chinese translations of the same book, the results identified two brain regions in the temporal lobe. BOLD signal in these regions was explained best by a model that acknowledges distinct subject and nonsubject grammatical roles, and embraces the ACT-R memory theory (van Rij et al. 2013). This cognitive model outperformed neural network models at the task of accounting for the observed BOLD variation. Subsequent work on pro-drop sentences, where the pronoun is not even mentioned, implicates the same brain regions (see Zhang et al. 2021, discussed above in section 4).

6. Lessons and next steps

The importance of interpretability The research highlighted in this review reaches for the goal of scientific explanation using computational models that are by-and-large interpretable. That is, the models connect with or implement theoretical constructs from linguistics.[8]

An alternative approach is suggested by the advent of high-performing deep neural networks that are trained end-to-end on specific natural language processing tasks such as language modeling (see RNN and Transformer entries in box 2). The success of such models relies upon their ability to extract fine-grained statistical regularities from exceptionally

[8]The conditions under which a language processing mechanism is interpretable with respect to linguistic theory have been debated. Stabler (1983) rules out a naïve box-ology. However the hypothesis under which a generative grammar is "computed but not represented" (Stabler's "Hh") is still viable. For a brief summary of this debate see Hale (2017, §3.4).

large training corpora. Schrimpf et al. (2020) use the internal states that result from such training to decode fMRI and electrophysiological data (see section 4). They report highly reliable fits between the neural network representations and brain data, especially in the left temporal lobe. The best-fitting models in their comparison are also the largest neural networks that show the highest performance on non-brain related measures.

While remarkable, the conclusions that can be drawn from such a comparison are theoretically limited. Like the brain itself, the neural networks investigated are "black boxes" in the sense mentioned above in section 5; their internal states are not directly interpretable The approach faces the pitfall of yielding trivial findings: systems that are highly sensitive to language statistics are in turn statistically similar to each other. Without interpretability, strong statistical relationships cannot form the foundation of explanatory theories. Examples of the latter can be found in related approaches, of course, as already seen with the attention payed to outputs of black-box models by Heilbron et al. (2021) and Caucheteux et al. (2021), and the interpretable feature-spaces of Wehbe et al. (2014a). Another path forward is to unpack the internal representations of neural networks (e.g. Yu & Ettinger 2020; Kuncoro et al. 2017).

Model spaces In seeking explanatory models, this review has drawn on studies whose basic mode of operation is to identify correlations between properties of a computational model and brain activity signals. When such models are interpretable, reliable correlations are taken as evidence that particular features of the model underlie the neural system being measured. But, this inferential step requires careful scrutiny. As already mentioned in section 4 above, statistical classifiers and artificial neural networks may latch on to idiosyncratic data patterns rather than linguistic features of interest (Heaven 2019 reviews this challenge as it faces modern AI more broadly.)

One strategy to strengthen inferencing relies on comparing alternative models of a particular linguistic phenomenon in terms of their statistical fit to neural data. This was done, for example, by Shain et al. 2020 for word-prediction, Anderson et al. 2021 for decoding structure, and Hale et al. 2018 for process models. In all of these examples, claims for or against a particular theoretical model are based on explicit comparison to a set of alternatives which together define a space of possible models or hypotheses. The logic of such comparisons is familiar from experimental design, where baselines are established with control conditions, or from linguistic analyses guided by evaluating minimal pairs, and all such efforts face familiar challenges (e.g. Cronbach & Meehl's Construct Validity).

In principle, such a model space ought to span all possible hypotheses that are plausible given current understanding of some research question. In practice it is more likely that the models span only a small part of the relevant space and inferences based on such a comparison are accordingly limited. To give just one concrete example, while Brennan et al. (2020) argue that neural indices of prediction reflect hierarchy, not just word sequences, that inference is based on comparing just three models. An alternative word-sequence model, could, in principle, yield different results for such a comparison. As in other domains, more general conclusions follow when results from multiple approaches converge; in this case, Shain et al. (2020) reach a theoretically similar conclusion from comparing a different set of sequence-based and hierarchical models. An opportunity granted by broad coverage computational models of naturalistic data is that researchers may compare alternative models against commonly shared datasets.

Table 1 Openly available neurolinguistic datasets for use with neuro-computational models.

Name	Description	Citation
The Little Prince Datasets	fMRI data from English (N=51), Chinese (N=35), and French-speaking (N=30) adults listening to a 1.5 h audiobook recording.	this paper; Franzluebbers et al. (2021)
The Alice Datasets	fMRI (N=26) and EEG (N=49) data from English-speaking adults listening to a 12 m audiobook recording.	Bhattasali et al. (2020)
Mother of all Unification Studies	fMRI and MEG data from Dutch-speaking participants who listened to (N=102) or read (N=102) a variety of sentences.	Schoffelen et al. (2019)
Naturalistic Neuroimaging Database	fMRI datasets from English-speaking adults (N=86) watching a full-length movie.	Aliko et al. (2020)
Narratives	fMRI datasets from English-speaking adults (N=345) who listened to a variety of spoken narratives.	Nastase et al. (2020)
Zurich Cognitive Language Processing Corpus	EEG datasets from English-speaking adults (N=12) who read 4–6 hours of natural text.	Hollenstein et al. (2020)

Reusable data With a view toward encouraging further progress, we invite the participation of linguists, neurobiologists and AI researchers by sharing several datasets of our own. These include EEG and fMRI data evoked by the first chapter of *Alice's Adventures in Wonderland* (Bhattasali et al. 2020), and fMRI data evoked by presentation of *The Little Prince* to native speakers of French, English and Chinese (Franzluebbers et al. 2021). These latter data are described in more detail in supplementary section 7. These datasets join a growing set of neurolinguistic corpora, described in Table 1, which offer neural data along side rich, often naturalistic, linguistic stimuli and annotations.

Progress on the Mapping Problem In 2012 David Poeppel dramatically declared that "there is absolutely no mapping to date that we understand in even the most vague sense" (page 50). A decade later we can be more optimistic. Now, quantitative linking hypotheses based on expectancy (section 3), decoding (section 4) and step-counting (section 5) allow us to test more granular brain/language correlations. In our judgment Poeppel's Radical Decomposition strategy has been productive. When we decompose a well-specified computational model into smaller pieces, each of which would apply hundreds or thousands of times in the course of a naturalistic listening session, it becomes possible to quantitatively compare model variants that are based on different leading ideas from linguistics. A technical prerequisite is that the models being compared must all be sufficiently broad in coverage to actually apply to the experimental stimuli to which human participants are exposed in a study. And, as we have argued above, it is important that the models being compared are linguistically-interpretable.

In the future, we anticipate the field moving beyond single-language studies to larger and larger sets of languages (e.g. as envisaged by Stehwien et al. 2020; Bornkessel-Schlesewsky

& Schlesewsky 2013). Simultaneously considering multiple languages in a single neurocomputational model will be crucial for drawing general conclusions. With the availability of web-scale corpora, such as crowdsourced encyclopedias, we believe this is goal is now within reach.

Disclosure Statement

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

Acknowledgments

This material is based upon work supported by the National Science Foundation under grant numbers 1903783 and 1607251, and the Jeffrey Sean Lehman Fund for Scholarly Exchange with China at Cornell University. The authors would like to thank Berta Franzluebbers for her assistance preparing the OpenNeuro archive.

7. Supplementary Information: Multilingual fMRI dataset

This section describes fMRI data elicited by the original French as well as English and Chinese translations of Saint-Exupéry's 1943 book *The Little Prince*. Upon acceptance of this article, the data will be publically-available at `https://openneuro.org/datasets/ds003643`.

7.1. Participants

Participants for the English study were 51 young adults (32 females, mean age=21.3, SD=3.6) with no history of psychiatric, neurological or other medical illness that might compromise cognitive functions. They self-identified as native English speakers, and strictly qualified as right-handed on the Edinburgh handedness inventory (Oldfield 1971). All participants were paid, and gave written informed consent prior to participation, in accordance with the IRB guidelines of Cornell University.

Chinese participants were 35 healthy, right-handed young adults (15 females, mean age=19.3, SD=1.6). They self-identified as native Chinese speakers, and had no history of psychiatric, neurological, or other medical illness that could compromise cognitive functions. All participants were paid, and gave written informed consent prior to participation, in accordance with the IRB guidelines of Jiangsu Normal University.

French participants were 30 healthy, right-handed adults (16 Female, mean age=24.3; SD=4.9). They were all native French speakers and had no history of psychiatric, neurological, or other medical illness that could compromise cognitive functions. All participants gave written informed consent prior to participation, in accordance with the Regional Committee for the Protection 195 of Persons involved in Biomedical Research.

7.2. Experimental Procedure

After giving their informed consent, participants were familiarized with the MRI facility and assumed a supine position on the scanner. Auditory stimuli were delivered through MRI-safe, high-fidelity headphones inside the head coil. The headphones were secured

against the plastic frame of the coil using foam blocks. An experimenter increased the sound volume stepwise until the participants could hear clearly. The stimuli were divided into 9 sections, and each lasted for about 10 minutes. Participants listened passively to the 9 sections and completed 4 quiz questions after each section (36 questions in total). These questions were used to confirm their comprehension and were viewed by the participants via a mirror attached to the head coil and they answered through a button box. The entire session, including preparation time and practice, lasted for around 2.5 hours.

7.3. Data Preprocessing

English and Chinese MRI images were acquired with a 3T MRI GE Discovery MR750 scanner with a 32-channel head coil. French MRI images were acquired with a 3T Siemens Magnetom Prisma Fit 230 scanner. Anatomical scans were acquired using a T1-weighted volumetric Magnetization Prepared RApid Gradient-Echo (MP-RAGE) pulse sequence. Functional scans were acquired using a multi-echo planar imaging (ME-EPI) sequence with online reconstruction (TR=2000 ms; English and Chinese: TEs=12.8, 27.5, 43 ms; French: TEs=10, 25, 38 ms; FA=77°; matrix size=72 x 72; FOV=240.0 mm x 240.0 mm; 2 x image acceleration; English and Chinese: 33 axial slices; French: 34 axial slices; voxel size=3.75 x 3.75 x 3.8 mm). All fMRI data were preprocessed using AFNI version 16 (Cox 1996). The first 4 volumes in each run were excluded from analyses to allow for T1-equilibration effects. Multi-echo independent components analysis (ME-ICA) (Kundu et al. 2012) were used to denoise data for motion, physiology and scanner artifacts. Images were then spatially normalized to the standard space of the Montreal Neurological Institute (MNI) atlas, yielding a volumetric time series resampled at 2 mm cubic voxels for the English and Chinese data and 3.15 mm cubic voxels for the French data.

LITERATURE CITED

Abney S, Johnson M. 1991. Memory requirements and local ambiguities of parsing strategies. *Journal of Psycholinguistic Research* 20:233–250

Aliko S, Huang J, Gheorghiu F, Meliss S, Skipper JI. 2020. A naturalistic neuroimaging database for understanding the brain using ecological stimuli. *Scientific Data* 7:347

Allen K, Pereira F, Botvinick M, Goldberg AE. 2012. Distinguishing grammatical constructions with fMRI pattern analysis. *Brain and Language* 123:174–182

Anderson AJ, Kiela D, Binder JR, Fernandino L, Humphries CJ, et al. 2021. Deep artificial neural networks reveal a distributed cortical network encoding propositional sentence-level meaning. *Journal of Neuroscience* 41:4100–4119

Baroni M. 2021. On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. Tech. rep., Facebook AI Research and Universitat Pompeu Fabra. ArXiv:2106.08694 or `https://ling.auf.net/lingbuzz/006031`

Baroni M, Bernardi R, Zamparelli R. 2014. Frege in space: A program for composition distributional semantics, In *Linguistic Issues in Language Technology, Volume 9, 2014 - Perspectives on Semantic Representations for Textual Inference*. CSLI Publications

Bastiaansen M, Hagoort P. 2006. Oscillatory neuronal dynamics during language comprehension. *Progress in Brain Research* 159:179–196

Bever TG. 1970. The cognitive basis for linguistic structures. In *Cognition and the Development of Language*, ed. J Hayes. New York: Wiley, 279–362

Bhattasali S, Brennan J, Luh WM, Franzluebbers B, Hale J. 2020. The Alice Datasets: fMRI & EEG observations of natural language comprehension, In *Proceedings of the 12th Language*

Resources and Evaluation Conference, pp. 120–125, Marseille, France: European Language Resources Association

Bhattasali S, Fabre M, Luh WM, Al Saied H, Constant M, et al. 2019. Localising memory retrieval and syntactic composition: An fMRI study of naturalistic language comprehension. *Language, Cognition and Neuroscience* 34:491–510

Blache P. 2018. Light-and-deep parsing: A cognitive model of sentence processing. In *Language, Cognition and Computational Models*, eds. T Poibeau, A Villavicencio. Cambridge, U.K.: Cambridge University Press, 27–52

Bornkessel-Schlesewsky I, Schlesewsky M. 2013. In Sanz et al. (2013), 241–252

Bornkessel-Schlesewsky I, Schlesewsky M. 2016. The importance of linguistic typology for the neurobiology of language. *Linguistic Typology* 20:303

Brennan J, Nir Y, Hasson U, Malach R, Heeger DJ, Pylkkänen L. 2012. Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and Language* 120:163–173

Brennan JR. 2016. Naturalistic sentence comprehension in the brain: Naturalistic comprehension. *Language and Linguistics Compass* 10:299–313

Brennan JR. 2020. Hemodynamic methods. In *Oxford Handbook of Experimental Syntax*, ed. J Sprouse. Oxford University Press

Brennan JR, Dyer C, Kuncoro A, Hale JT. 2020. Localizing syntactic predictions using recurrent neural network grammars. *Neuropsychologia* 146:107479

Brennan JR, Hale JT. 2019. Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PLOS ONE* 14:e0207741

Brennan JR, Pylkkänen L. 2017. MEG evidence for incremental sentence composition in the anterior temporal lobe. *Cognitive Science* 41:1515–1531

Brennan JR, Stabler EP, Van Wagenen SE, Luh WM, Hale JT. 2016. Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and Language* 157-158:81–94

Brouwer H, Delogu F, Venhuizen NJ, Crocker MW. 2021. Neurobehavioral correlates of surprisal in language comprehension: A neurocomputational model. *Frontiers in Psychology* 12:110

Buchweitz A, Shinkareva SV, Mason RA, Mitchell TM, Just MA. 2012. Identifying bilingual semantic neural representations across languages. *Brain and Language* 120:282 – 289

Caplan D. 1992. Language: structure, processing, and disorders. MIT Press

Caucheteux C, Gramfort A, King JR. 2021. Disentangling syntax and semantics in the brain with deep networks, In *Proceedings of the 38th International Conference on Machine Learning*, eds. M Meila, T Zhang, vol. 139 of *Proceedings of Machine Learning Research*, pp. 1336–1348, PMLR

Choe DK, Charniak E. 2016. Parsing as language modeling, In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2331–2336, Austin, Texas: Association for Computational Linguistics

Chomsky N. 1956. Three models for the description of language. *IRE Transactions on Information Theory* 2:113–124

Coecke B, Sadrzadeh M, Clark S. 2010. Mathematical foundations for a compositional distributional model of meaning. *Linguistic Analysis* 36:345–384

Correia J, Formisano E, Valente G, Hausfeld L, Jansma B, Bonte M. 2013. Brain-based translation: fMRI decoding of spoken words in bilinguals reveals language-independent semantic representations in anterior temporal lobe. *The Journal of Neuroscience* 34:332

Cox RW. 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research, an International Journal* 29:162–173

Crinion J, Turner R, Grogan A, Hanakawa T, Noppeney U, et al. 2006. Language control in the bilingual brain. *Science* 312:1537–1540

Cronbach LJ, Meehl PE. 1955. Construct validity in psychological tests. *Psychological Bulletin* 52:281–302

de Saint-Exupéry A. 1943. Le petit prince. Harcout, Brace and World

Dikker S, Pylkkänen L. 2013. Predicting language: MEG evidence for lexical preactivation. *Brain and Language* 127:55–64

Dunagan D, Zhang S, Li J, Bhattasali S, Pallier C, et al. 2021. Neural correlates of semantic number: A cross-linguistic investigation. doi:10.1101/2021.05.11.443670

Dyer C, Kuncoro A, Ballesteros M, Smith NA. 2016. Recurrent neural network grammars, In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 199–209, San Diego, California: Association for Computational Linguistics

Eisenstein J. 2019. Introduction to natural language processing. MIT Press

Elman J. 1990. Finding structure in time. *Cognitive Science* 14:179–211

Embick D, Poeppel D. 2015. Towards a computational(ist) neurobiology of language: Correlational, integrated, and explanatory neurolinguistics. *Language, cognition and neuroscience* 30:357–366

Ettinger A, Feldman N, Resnik P, Phillips C. 2016. Modeling N400 amplitude using vector space models of word representation. In *Proceedings of the 38^{th} Annual Meeting of the Cognitive Science Society*, eds. A Papafragou, D Grodner, D Mirman, J Trueswell. 1445–1450

Evert S. 2009. Corpora and collocations. In *Corpus linguistics: An international handbook*, eds. A Lüdeling, M Kytö, vol. 2. Berlin, Germany: de Gruyter, 1212–1248

Fillmore CJ. 1968. The case for case. In *Universals in linguistic theory*, eds. EW Bach, RT Harms. Holt, Rinehart and Winston, 1–88

Frank SL, Otten LJ, Galli G, Vigliocco G. 2015. The ERP response to the amount of information conveyed by words in sentences. *Brain and Language* 140:1–11

Frankland SM, Greene JD. 2015. An architecture for encoding sentence meaning in left mid-superior temporal cortex. *Proceedings of the National Academy of Sciences* 112:11732–11737

Franzluebbers B, et al. 2021. The Little Prince fMRI dataset: English, French & Chinese. OpenNeuro. Dataset ds003643

Frazier L. 1985. Syntactic complexity. In *Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives*, eds. D Dowty, L Karttunen, AM Zwicky. Cambridge Univ Press, 129–187

Friederici AD. 2017. Language in our brain: The origins of a uniquely human capacity. MIT Press

Friston K. 2010. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience* 11:127–138

Goldberg Y. 2017. Neural network methods in natural language processing. Synthesis Lectures on Human Language Technologies: Lecture number 37. Morgan & Claypool

Hagoort P. 2016. MUC (memory, unification, control): A model on the neurobiology of language beyond single word processing. In *Neurobiology of Language*, eds. G Hickok, SL Small. Academic Press

Hale J. 2016. Information-theoretical complexity metrics. *Language and Linguistics Compass* 10:397–412

Hale J, Kuncoro A, Hall K, Dyer C, Brennan J. 2019. Text genre and training data size in human-like parsing, In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5846–5852, Hong Kong, China: Association for Computational Linguistics

Hale JT. 2017. Models of human sentence comprehension in computational psycholinguistics. In *Oxford Research Encyclopedia of Linguistics*, ed. M Aronoff. Oxford University Press

Hale JT, Dyer C, Kuncoro A, Brennan JR. 2018. Finding syntax in human encephalography with beam search, In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2727–2736, Melbourne, Australia: Association for Computational Linguistics

Hansen PC, Kringelbach ML, Salmelin R. 2010. MEG: an introduction to methods. Oxford University Press

Hashemzadeh M, Kaufeld G, White M, Martin AE, Fyshe A. 2020. From language to language-ish:

How brain-like is an LSTM's representation of nonsensical language stimuli?, In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 645–656, Online: Association for Computational Linguistics

Heaven D. 2019. Why deep-learning AIs are so easy to fool. *Nature* 574:163–166

Heilbron M, Armeni K, Schoffelen JM, Hagoort P, de Lange FP. 2021. A hierarchy of linguistic predictions during natural language comprehension. *bioRxiv*

Hemforth B, Konieczny L. 2006. Language processing: Construction of mental models or more? In *Mental Models and the Mind*, eds. C Held, M Knauff, G Vosgerau, vol. 138 of *Advances in Psychology*. North-Holland, 189–204

Henderson JM, Choi W, Lowder MW, Ferreira F. 2016. Language structure in the brain: A fixation-related fMRI study of syntactic surprisal in reading. *NeuroImage* 132:293–300

Hochreiter S, Schmidhuber J. 1997. Long Short-Term Memory. *Neural Computation* 9:1735–1780

Hollenstein N, Troendle M, Zhang C, Langer N. 2020. ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation, In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 138–146, Marseille, France: European Language Resources Association

Honey CJ, Thompson CR, Lerner Y, Hasson U. 2012. Not Lost in Translation: Neural Responses Shared Across Languages. *Journal of Neuroscience* 32:15277–15283

Huth AG, de Heer WA, Griffiths TL, Theunissen FE, Gallant JL. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* 532:453–458

Jelinek F, Lafferty JD. 1991. Computation of the probability of initial substring generation by stochastic context-free grammars. *Computational Linguistics* 17

Joshi A. 1985. How much context-sensitivity is required to provide reasonable structural descriptions: Tree adjoining grammars. In *Natural language parsing: Psychological, computational, and theoretical perspectives*, eds. D Dowty, L Karttunen, A Zwicky. Cambridge: Cambridge University Press, 206–250

Joshi AK, Vijay-Shanker K, Weir D. 1991. The convergence of mildly context-sensitive grammatical formalisms, In *Foundational Issues in Natural Language Processing*, pp. 31–81, MIT Press

Jurafsky D, Martin JH. 2021. Speech and language processing. Prentice-Hall, 3rd ed.

Kallmeyer L. 2010. Parsing beyond context-free grammars. Springer

Kaplan RM. 1972. Augmented transition networks as psychological models of sentence comprehension. *Artificial Intelligence* 3:77–100

Katz JJ, Fodor JA. 1963. The structure of a semantic theory. *Language* 39:170–210

Kemmerer D. 2014. Cognitive neuroscience of language. Psychology Press, New York

Kriegeskorte N, Goebel R, Bandettini P. 2006. Information-based functional brain mapping. *Proceedings of the National Academy of Sciences* 103:3863–3868

Kuncoro A, Ballesteros M, Kong L, Dyer C, Neubig G, Smith NA. 2017. What do recurrent neural network grammars learn about syntax?, In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 1249–1258, Valencia, Spain: Association for Computational Linguistics

Kundu P, Inati SJ, Evans JW, Luh WM, Bandettini PA. 2012. Differentiating {BOLD} and non-bold signals in fmri time series using multi-echo {EPI}. *NeuroImage* 60:1759 – 1770

Kutas M, Federmeier KD. 2000. Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Science* 4:463–469

Langendoen DT. 2008. Coordinate grammar. *Language* 84:691–709

Lenneberg EH. 1967. Biological foundations of language. John Wiley & Sons

Lerner Y, Honey CJ, Silbert LJ, Hasson U. 2011. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience* 31:2906–2915

Li J, Brennan J, Mahar A, Hale J. 2016. Temporal lobes as combinatory engines for both form and meaning, In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity*, pp. 186–191

Li J, Hale JT. 2019. Grammatical predictors for fMRI timecourses. In *Minimalist parsing*, eds. EP Stabler, RC Berwick. New York, NY: Oxford University Press, 159–173

Li J, Wang S, Luh WM, Pylkkänen L, Yang Y, Hale JT. 2020. Modeling pronoun resolution in the brain. *bioRxiv*

Linzen T, Baroni M. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics* 7:195–212

Linzen T, Dupoux E, Goldberg Y. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics* 4:521–535

Lopopolo A, Frank SL, van den Bosch A, Willems RM. 2017. Using stochastic language models (SLM) to map lexical, syntactic, and phonological information processing in the brain. *PLOS ONE* 12:e0177794

Luck SJ. 2014. An introduction to the event-related potential technique. Cambridge, Massachusetts: MIT Press, 2nd ed.

Lyu B, Choi HS, Marslen-Wilson WD, Clarke A, Randall B, Tyler LK. 2019. Neural dynamics of semantic composition. *Proceedings of the National Academy of Sciences* 116:21318–21327

Manning CD, Schütze H. 2000. Foundations of statistical natural language processing. MIT Press

Marcus MP, Santorini B, Marcinkiewicz MA. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19:313–330

Marslen-Wilson WD. 1975. Sentence perception as an interactive parallel process. *Science* 189:226–228

Martin AE, Doumas LAA. 2017. A mechanism for the cortical computation of hierarchical linguistic structure. *PLoS biology* 15:e2000663

Martin AE, Doumas LAA. 2019. Tensors and compositionality in neural systems. *Philosophical Transactions B* 375:20190306

Matchin W, Hammerly C, Lau EF. 2016. The role of the IFG and pSTS in syntactic prediction: Evidence from a parametric study of hierarchical structure in fMRI. *Cortex; a journal devoted to the study of the nervous system and behavior* 88:106–123

McClelland JL, Kawamoto AH. 1986. In Rumelhart et al. (1986b), 272–325

Meyer L, Sun Y, Martin AE. 2020. Synchronous, but not entrained: Exogenous and endogenous cortical rhythms of speech and language processing. *Language, Cognition and Neuroscience* 35:1089–1099

Mitchell J, Lapata M. 2010. Composition in distributional models of semantics. *Cognitive Science* 34:1388–1429

Mitchell TM, Shinkareva SV, Carlson A, Chang KM, Malave VL, et al. 2008. Predicting human brain activity associated with the meanings of nouns. *Science* 320:1191–1195

Murphy B, Wehbe L, Fyshe A. 2018. Decoding language from the brain. In *Language, cognition, and computational models*, eds. T Poibeau, A Villavicencio. Cambridge, U.K.: Cambridge University Press, 53–80

Nastase SA, Liu YF, Hillman H, Zadbood A, Hasenfratz L, et al. 2020. Narratives: fMRI data for evaluating models of naturalistic language comprehension. Preprint, Neuroscience

Nelson MJ, El Karoui I, Giber K, Yang X, Cohen L, et al. 2017. Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences of the United States of America* 114:E3669–E3678

O'Donnell TJ. 2015. Productivity and reuse in language: A theory of linguistic computation and storage. The MIT Press

Oldfield RC. 1971. The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia* 9:97–113

Pallier C, Devauchelle AD, Dehaene S. 2011. Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences* 108:2522–2527

Partee BH, ter Meulen A, Wall RE. 1993. Mathematical methods in linguistics. Kluwer

Pereira F, Lou B, Pritchett B, Ritter S, Gershman SJ, et al. 2018. Toward a universal decoder of

linguistic meaning from brain activation. *Nature communications* 9:963

Pereira F, Mitchell T, Botvinick M. 2009. Machine learning classifiers and fMRI: A tutorial overview. *NeuroImage* 45:S199–S209

Pesetsky D. 1995. Zero syntax: Experiencers and cascades. Cambridge, MA: MIT Press

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

Pylkkänen L. 2019. The neural basis of combinatory syntax and semantics. *Science* 366:62–66

Qian P, Qiu X, Huang X. 2016. Bridging LSTM architecture and the neural dynamics during reading, In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pp. 1953–1959

Rabovsky M, Hansen SS, McClelland JL. 2018. Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour* 2:693–705

Reddy AJ, Wehbe L. 2021. Can fMRI reveal the representation of syntactic structure in the brain? *Under review*

Rescorla M. 2020. The Computational Theory of Mind. The Stanford Encyclopedia of Philosophy

Rohde DL. 2002. A connectionist model of sentence comprehension and production. Ph.D. thesis, Carnegie Mellon University

Rumelhart DE, Hinton GE, McClelland JL. 1986a. In Rumelhart et al. (1986b), chap. 8, 318–362

Rumelhart DE, McClelland J, the PDP Research Group. 1986b. Parallel distributed processing: Explorations in the microstructure of cognition. Cambridge, MA: MIT Press

Salmelin R, Baillet S. 2009. Electromagnetic brain imaging. *Human brain mapping* 30:1753–7

Sanz M, Laka I, Tanenhaus MK, eds. 2013. Language down the garden path: The cognitive and biological basis of linguistic structures. Oxford University Press

Schoffelen JM, Oostenveld R, Lam NHL, Uddén J, Hultén A, Hagoort P. 2019. A 204-subject multimodal neuroimaging dataset to study language processing. *Scientific Data* 6:17

Schrimpf M, Blank I, Tuckute G, Kauf C, Hosseini EA, et al. 2020. The neural architecture of language: Integrative reverse-engineering converges on a model for predictive processing. *bioRxiv* :2020.06.26.174482

Shain C, Blank IA, van Schijndel M, Schuler W, Fedorenko E. 2020. fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia* 138:107307. Data available at `doi:10.17605/osf.io/eyp8q`

Smith GT. 2005. On construct validity: Issues of method and measurement. *Psychological Assessment* 17:396–408

Smolensky P. 2015. Four facts about Tensor Product Representations. talk given at NIPS workshop: Cognitive Computation: Integrating Neural and Symbolic Approaches. `https://youtu.be/teuJ4SngxjQ`

Smolensky P, Legendre G. 2006. The Harmonic Mind. MIT Press

Stabler E. 2001. Minimalist grammars and recognition. In *Linguistic form and its computation*, chap. 10. CSLI publications, 327–352

Stabler EP. 2013. In Sanz et al. (2013), 316–323

Stabler Jr. EP. 1983. How are grammars represented? *Behavioral and Brain Sciences* 6:391–421

Stanojević M, Bhattasali S, Dunagan D, Campanelli L, Steedman M, Hale J. 2021. Modeling incremental language comprehension in the brain with Combinatory Categorial Grammar, In *Proceedings of the Computational Linguistics and Cognitive Modeling workshop*. To Appear

Steedman M. 1999. Connectionist sentence processing in perspective. *Cognitive Science* 23:615–634

Steedman M. 2000. The syntactic process. Cambridge, MA: MIT Press

Stehwien S, Henke L, Hale J, Brennan J, Meyer L. 2020. The little prince in 26 languages: Towards a multilingual neuro-cognitive corpus, In *Proceedings of the Second Workshop on Linguistic and Neurocognitive Resources*, pp. 43–49, Marseille, France: European Language Resources Association

Stowe LA, Broere CA, Paans AM, Wijers AA, Mulder G, et al. 1998. Localizing components of a

complex task: Sentence processing and working memory. *Neuroreport* 9:2995–2999

Stowe LA, Haverkort M, Zwarts F. 2005. Rethinking the neurological basis of language. *Lingua* 115:997–1042

Swaab TY, Ledoux K, Camblin CC, Boudewyn MA. 2012. Language-related ERP components. In *The Oxford Handbook of Event-Related Potential Components*, eds. SJ Luck, ES Kappenman. Oxford University Press, 397–439

Toneva M, Wehbe L. 2019. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems 32*, eds. H Wallach, H Larochelle, A Beygelzimer, F d'Alché Buc, E Fox, R Garnett. Curran Associates, Inc., 14954–14964

Townsend DJ, Bever TG. 2001. Sentence comprehension: The integration of habits and rules. MIT Press

Ullman MT. 2004. Contributions of memory circuits to language: The declarative/procedural model. *Cognition* 92:231–70

van Rij J, van Rijn H, Hendriks P. 2013. How wm load influences linguistic processing in adults: A computational model of pronoun interpretation in discourse. *Topics in Cognitive Science* 5:564–580

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al. 2017. Attention is all you need, In *Proceedings of Neural Information Processing Systems*, pp. 6000–6010

Vinyals O, Kaiser Lu, Koo T, Petrov S, Sutskever I, Hinton G. 2015. Grammar as a foreign language, In *Advances in Neural Information Processing Systems*, eds. C Cortes, N Lawrence, D Lee, M Sugiyama, R Garnett, vol. 28. Curran Associates, Inc.

Wedekind J, Kaplan RM. 2020. Tractable Lexical-Functional Grammar. *Computational Linguistics* 46:515–569

Wehbe L, Murphy B, Talukdar P, Fyshe A, Ramdas A, Mitchell T. 2014a. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PLoS ONE* 9:e112575

Wehbe L, Vaswani A, Knight K, Mitchell T. 2014b. Aligning context-based statistical models of language with brain activity during reading, In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 233–243, Doha, Qatar: Association for Computational Linguistics

Werbos PJ. 1990. Backpropagation through time: What it does and how to do it. *Proceedings of the IEEE* 78:1550–1560

Willems RM, Frank SL, Nijhof AD, Hagoort P, van den Bosch A. 2015. Prediction During Natural Language Comprehension. *Cerebral Cortex* 26:2506–2516

Yarkoni T, Speer NK, Balota DA, McAvoy MP, Zacks JM. 2008. Pictures of a thousand words: Investigating the neural mechanisms of reading with extremely rapid event-related fMRI. *NeuroImage* 42:973–987

Yu L, Ettinger A. 2020. Assessing Phrasal Representation and Composition in Transformers, In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4896–4907, Online: Association for Computational Linguistics

Zhang S, Li J, Yang Y, , Hale J. 2021. Decoding the silence: neural bases of zero pronoun resolution in Chinese. doi:10.1101/2021.05.06.442989