Simulated Language Acquisition in a Biologically Realistic Model of the Brain

Daniel Mitropolsky
McGovern Institute for Brain Research
MIT
Cambridge, MA
mitropol@mit.edu

Christos H. Papadimitriou
Department of Computer Science
Columbia University
New York, NY
christos@columbia.edu

Abstract

Despite tremendous progress in neuroscience, we do not have a compelling narrative for the precise way whereby the spiking of neurons in our brain results in high-level cognitive phenomena such as planning and language. We introduce a simple mathematical formulation of six basic and broadly accepted principles of neuroscience: excitatory neurons, brain areas, random synapses, Hebbian plasticity, local inhibition, and inter-area inhibition. We implement a simulated neuromorphic system based on this formalism, which is capable of basic language acquisition: Starting from a tabula rasa, the system learns, in any language, the semantics of words, their syntactic role (verb versus noun), and the word order of the language, including the ability to generate novel sentences, through the exposure to a modest number of grounded sentences in the same language. We discuss several possible extensions and implications of this result.

1 Introduction

It is beyond doubt that cognitive phenomena such as language, reasoning, and planning are the direct product of the activity of neurons and synapses, and yet there is no existing overarching theory that explains exactly how this could be done. In the words of Richard Axel [1], *"we do not have a logic for the transformation of neural activity into thought and action. I view discerning [this] logic as the most important future direction of neuroscience."* Through the use of the term "logic" in this statement, Axel seems to suggest that the research community should strive to identify, from among the ever growing volume of neuroscientific knowledge, see e.g. [21], a conceptual core of basic elements and principles which suffice for bringing about the brain's function.

We introduce such a minimalistic neural model of the brain, which we call NEMO (see Figure 1, and [32]), a simple mathematical formulation of six basic and uncontroversial ideas in neuroscience: excitatory neurons, brain areas, random synaptic connectivity, local inhibition in each area, Hebbian plasticity, and interarea inhibition. Importantly, NEMO can be simulated very efficiently at the scale of tens of millions of neurons and trillions of synapses. To test whether NEMO qualifies as the abstraction sought in Axel's statement, we design a NEMO system that can carry out what is arguably the most prominent achievement of the animal brain, namely *natural language acquisition.* In past work on NEMO it has been shown, through both mathematical proofs and simulations, that it is capable of carrying out basic computations through the formation and manipulation of neural assemblies [32], as well as various tasks of a cognitive nature [7, 12], including *parsing* natural language sentences [26, 27]. However, that NEMO parser presupposed that the words in the language are already represented as assemblies of neurons whose synaptic connections encode the grammatical role of each word.

Preprint. Under review.

Here, we demonstrate that lexicon, syntax and semantics of a language can be *learned* by a tabula rasa NEMO system in a way that mirrors human language acquisition. In particular, our system learns (a) semantic representations of concrete nouns and verbs; (b) basic syntactic characteristics such as the part of speech for each word; (c) the language's word order, in several moods, and (d) to generate novel sentences — all through exposure to grounded language.

Language acquisition, of course, entails much more than this. In particular, a complete account must include the acquisition of speech sounds and word phonetics, parts-of-speech other than nouns and verbs, including functional categories of words such as "the," abstract words, and syntax beyond basic word-order. Furthermore, in the baby brain the learning of all these aspects is intertwined in subtle ways [11]. However, since our purpose here is to design a rigorous experiment that establishes that basic language acquisition can be carried out by a NEMO artifact, we focus on implementing the core statistical learning tasks (a)–(d) above in two phases: first the learning of word semantics and part of speech (a–b) and then the learning of word order and generation (c–d). As for speech sounds and phonetics, we bypass this phase by adopting an input-output convention whereby a sentence is presented to the system as a sequence of stimuli corresponding to word tokens, and similarly for sentence generation.

In our experiments, the learning device is exposed to *grounded whole sentences* in any fixed natural language (we use several languages in our experiments). By "grounded" we mean that each sentence is presented while representations corresponding to the sentence are active in the sensory and motor cortex. That is, when the sentence "the dog runs" is input, we assume that in the premotor cortex the mirror cells [23, 13] for "run," as well as the generic dog representation in the inferotemporal visual cortex [22], are firing.

In a very interesting precursor to this work, it was shown in [37] that concrete nouns, verbs and their meanings can be learned through grounded presentation in a different biologically realistic model of the brain. However, that work was about the easier problem of learning *isolated words,* without syntactic context — it is known that nonhuman primates can achieve this level of language learning, see, for example [36]. Here we solve the far more demanding problem — and an exclusive characteristic of humans — of learning word semantics and the verb vs. noun distinction, as well as word order, through exposure to grounded whole sentences. We believe that our results entail the first biologically plausible artifact that is capable of basic language acquisition; if scientists were to discover such linguistic competence in a nonhuman animal, we believe that this finding would fundamentally change the currently held view on the exclusivity of language.

Besides addressing the quest for Axel's "logic" mentioned above, the present work also pertains to another long-standing fundamental question, namely: *What is the biological basis of language?* In particular, is language the fruit of unique human characteristics at the genetic, molecular, and/or neural level? In the past, the discovery of such characteristics (e.g., the FOXP2 gene, for a brief time called "the language gene," and the novel dendritic potentials recently discovered in human layer 2-3 neurons [18]) has led to speculation about the biological roots of the human linguistic ability, see, e.g., [8]. The biologically plausible language acquisition system presented here can be seen as evidence for an alternative *null hypothesis:* A language system can be built on top of the mammalian brain, through simple and widely accepted neuroscientific elements and principles.

The specific neural architecture and algorithms we propose can be seen as a comprehensive, neurobiologically plausible *hypothesis* for the human language system, a domain in which such concrete hypotheses are scarce. The brain areas used in our model — their number, function and synaptic connections between them — are arguably compatible with existing neuroanatomical and psycholinguistic evidence and opinion, and reflect the field's consensus where it exists. In fact, our artificial language system can be the source of predictions about activations in these brain areas in the course of comprehension and generation, which may prove useful in neurolinguistic research.

Finally, let us note here that, compared to another extant class of computational artifacts with linguistic abilities, namely the large language models (LLMs) of now [29], our system has the important distinction of *biological plausibility* — in particular, it does not use backpropagation, the most important ingredient of current machine learning, which has never been observed in the animal brain.

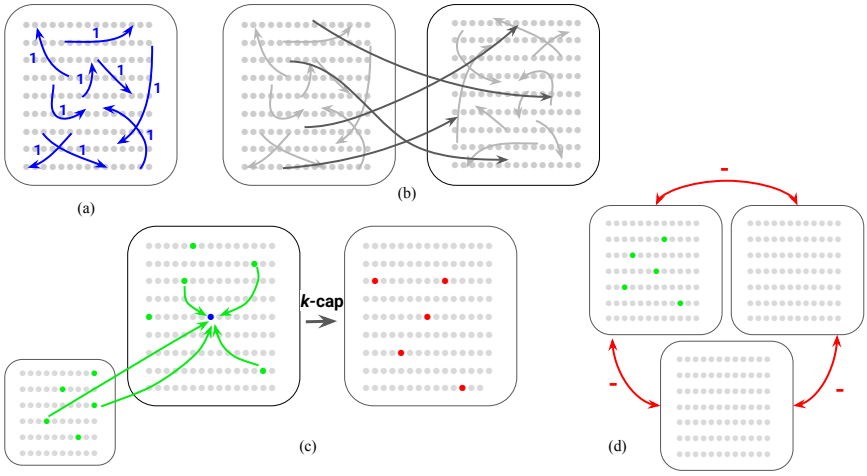

Figure 1: An overview of our biologically constrained neural model (NEMO for short). The brain is modeled as a set of brain *areas*, each of which (a) has n neurons; in our experiments n is typically a million. For every ordered pair (i, j) of neurons in an area, a directed synapse from i to j is included with probability p, initialized with weight 1, resulting in a directed random synaptic graph; in our experiments, p is typically equal to 0.01. (b) NEMO artifacts consist of multiple areas; each ordered pair of areas may be connected by a *fiber* in which case each possible synapse from each neuron in the first area to each neuron in the second is included with probability p. Only a few of all possible pairs of brain areas are connected by a fiber, defining the top level of the system's architecture. (c) NEMO is a dynamical system defined by the *firing* operation. Time proceeds in discrete steps. At each time step only k neurons in each area fire — a selection that models local inhibition in the area — where k is much smaller than n (in our experiments, k is typically about $\sqrt{n}$). Of all n neurons in an area the neurons that fire are those k that have received the *largest synaptic input* from the previous time-step. That is, for each neuron i the quantity $\sum_{\text{synapses } (j,i)} w_{j,i}$ is computed, where the sum is taken over all neurons j, in the same area or in other areas, that fired in the previous step and have a synapse to i, and $w_{j,i}$ stands for the weight of this synapse. Once these sums have been computed for all neurons in the area, the top k neurons are selected to fire in the next time step, with any ties resolved arbitrarily. This primitive of the NEMO model is called *k-cap*. In the illustration above, to the left of the "k-cap" arrow we show an area with firing neurons (green), as well as firing neurons in another area, and an example neuron (dark blue) showing its incoming edges from firing neurons (such edges from firing neurons to other neurons are not shown). To the right of the "k-cap" arrow is the state of the area after inputs have been summed and "k-cap" applied; the dark blue example neuron from the previous time step happened to be among the five top-k inputs, and is hence firing (red); in this example, $k = 5$. There is also *Hebbian plasticity:* for two synaptically connected neurons j, i firing in succession for two consecutive steps in this order, the synaptic weight $w_{j,i}$ is multiplied by $(1 + \beta)$; the positive plasticity parameter β is typically 0.05. (d) Additionally, two or more areas can be in *mutual inhibition*, in which case there is firing only in the area that receives the greatest total synaptic input from the previous step, and no firing in the other areas. The set of areas, fibers, mutually-inhibited sets of areas, as well as n, p, k, and β are the *hyperparameters* of the NEMO model. In our experiments, a local hyperparameter such as p and β may vary from one area or fiber to the other.

2 Procedures and Results

2.1 Learning word semantics

We start by describing how the NEMO system for language acquisition learns the meaning of concrete nouns and intransitive verbs, as well as their syntactic role (noun versus verb), see Figure 2, and see SI for the parameter values. In the next section we describe the extensions of this core system that are needed for learning syntax and the word order of the language (compare Figures 2 and 4).

The earliest part of language acquisition during roughly the first year of life involves learning the speech sounds of the language. In our experiment we bypass this stage by adopting a convenient *input/output convention:* Words are input to the device through the activation of word representations in an area that we call PHON, where phonetic knowledge for recognizing and articulating words is already in place. There seems to be a consensus in neurolinguistic literature (see, e.g., [19, 30]) that the human brain indeed contains such an area. PHON an *input area*, in the sense that it is initialized with assemblies, each representing the phonetics of a word in the language. This input-output convention is tantamount to assuming that the device knows *ab ovo* how to recognize and articulate the words in the lexicon, but at the start it is oblivious of their meaning or part of speech.

The *lexicon* consists of two areas, LEX_1 and LEX_2, which start as a tabula rasa, but in the experiment will end up forming one assembly representation for each word in the lexicon: in LEX_1 for nouns and in LEX_2 for verbs. This reflects a hypothesis in neurolonguistics, based on neurological studies and separation studies between nouns and verbs, that nouns and verbs are represented in different areas of the brain [28, 38, 24]. There are fibers that go, both ways, between PHON and each of LEX_1 and LEX_2.

The acquisition system also has a number of *semantic areas*, sensorimotor and other cortical areas that feed into the language system; see, e.g., [34] for results suggesting the existence of many such areas. Two of these areas, VISUAL and MOTOR, have special significance, but there are also several additional context areas representing other modes of semantic context such as olfactory, non-linguistic auditory, location, emotion, etc., and labeled C_i for $i = 1, \ldots, m$. Like PHON, the semantic areas are input areas, initialized to contain assembly representations that provide semantic content to words in the lexicon, with action verbs taking their meaning predominantly from mirror neurons in the motor cortex and concrete nouns from visual representations in the visual cortex. The exact total number m of extra context areas will vary in our experiment; our algorithm works for various m, from zero to ten.

There are fibers connecting the lexical and semantic areas — it is known that there are many evolutionarily novel fibers in the vicinity of the left medial temporal lobe where the lexicon is believed to reside in the human brain, see e.g. [14]. In our model there are fibers in both directions between LEX_1 and LEX_2 on the one side, and PHON and the semantic areas on the other. Four of these $2m + 6$ fibers, namely the ones between PHON and the two lexical areas, as well as the one between LEX_1 and VISUAL, and the one between LEX_2 and MOTOR, have increased parameters β and p, making them stronger conduits of synaptic input — these fibers are depicted as bolder arrows in the figure.

The word learning experiment. We fix a natural language and ℓ concrete nouns and ℓ intransitive action verbs, where ℓ is a parameter that we vary in our experiment. By restricting ourselves to action nouns and concrete verbs we assure that for each word in the lexicon there is a corresponding semantic representation in the semantic areas. The restriction to intransitive verbs is considered only for the present experiment of learning word semantics; transitive verbs will be introduced soon. All sentences in the experiment are of length two: "the cat jumps" and "the dog eats," where, for the purposes of our experiment, we assume that function words such as "the" are simply ignored; taking into account such words makes the learning task easier. Note that the language can have either SV (subject-verb) principal word order (as in English, Chinese and Swahili) or VS (as in Irish, Modern Standard Arabic, and Tagalog), and our model should succeed in either scenario; this is one of the main challenges of this experiment.

For each word w we initialize an assembly $\text{PHON}[w]$ in PHON. In addition, if w is a noun, we initialize an assembly in VISUAL corresponding to the visual representation of the object of w in the inferotemporal cortex, and if w is a verb, there is a corresponding assembly in MOTOR — presumably

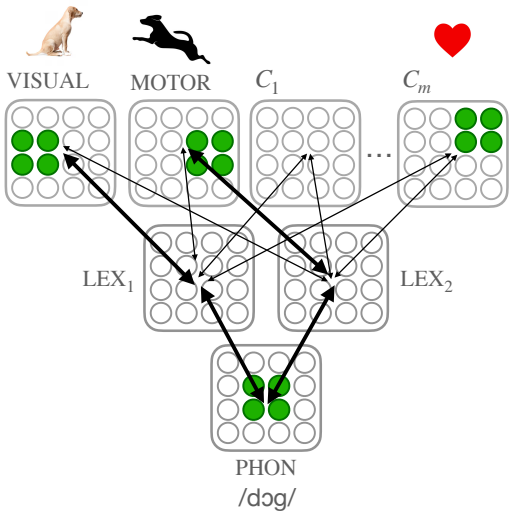

Figure 2: The NEMO system for learning concrete nouns and verbs. PHON is an input area which is initialized with one representation, for each word, of the phonological code of the word. The semantic areas VISUAL, MOTOR and $C_1, \ldots, C_m$ are also input areas, with visual representations, motor representations, and other perceptual representations (e.g. olfactory, emotional, etc). LEX_1 and LEX_2 are the areas where the nouns and verbs, respectively, in the language will be learned. LEX_1 has high connectivity to VISUAL and LEX_2 to MOTOR, and lower connectivity to the other semantic areas. The figure shows the firing pattern in the input areas when input the sentence "dog jumps," at the moment "dog" is heard.

the mirror cells for w in the premotor cortex. Finally, for a subset of the remaining m context areas, selected at random, we create in each area an assembly corresponding to w.

We input randomly generated sentences via a uniform choice of noun and verb. Note that this results in including bizarre sentences and impossible scenes, such as "the trees jump", but we have verified that omitting these does not affect the results. To input a sentence, we input each word individually, in the word order of the language, by firing the corresponding assembly in PHON τ times, where τ is a parameter of the experiment, while the contextual input of the scene (the assemblies associated with the words in the sentence, both with the noun and the verb) fire throughout the duration of the entire sentence. That is, as in life, for the sentence "the cat jumps," the VISUAL assembly representing "cat" and the MOTOR assembly representing "jump" both fire as the learner hears "cat" and then "jumps." Note that it is a priori impossible to know which of the two words is the concrete noun, and which is the action verb. However, as we shall see, the interaction between NEMO dynamics and cooccurrence statistics results in the system learning to correctly distinguish between noun and verb.

What does it mean for the system to succeed in learning these nouns and verbs? Informally, we would like LEX_1 to become a "noun area," and LEX_2 a "verb area," containing *robust neural representations of each noun and verb* in the lexicon, respectively, and we would also like each of these robust word representations to be somehow aware of the meaning of the corresponding word, as well as aware of their affinity with the corresponding phonetic assembly in PHON.

More formally, we require that, after the input of enough sentences, the following properties hold:

Property 1: Formation of word assemblies. For each noun, if we fire its representation in PHON, or that in VISUAL, or both, the resulting k-cap in LEX_1 is so densely interconnected that, if we fire it by itself, the k-cap in LEX_1 remains the same set of neurons again and again; and that this also occurs if we continue to fire the representation in VISUAL. In the same sense, a neural assembly emerges for each verb in LEX_2.

Property 2: Formation of the PHON-LEX-VISUAL pathway. The creation of assemblies for each word enables a pathway between PHON and the semantic areas via LEX_1 and LEX_2. The pathway is bidirectional: for each noun w, if we fire PHON$[w]$ for 2 time steps, via LEX_1 we activate VISUAL$[w]$ and its representation in the contextual areas. Conversely, if we fire VISUAL$[w]$ and the contextual representations, we retrieve PHON$[w]$. In the same way, a pathway is forged for verbs.

Property 3: Classification of Nouns vs. Verbs The final property required to emerge from learning is that, while very robust assemblies for nouns are formed in LEX_1 and for verbs in LEX_2, *nothing akin* to a stable assembly representation exists for nouns in LEX_2, nor for verbs in LEX_1. For a noun w, firing $\text{PHON}[w]$ into LEX_2 results in a "wobbly" set of neurons: firing this set again recurrently, *even with* the input of $\text{PHON}[w]$, results in a quite different set. Additionally, firing this set *back* into PHON results in a set that has very little overlap with $\text{PHON}[w]$. This last property yields an algorithmic test for whether a word is a noun or verb: try firing $\text{PHON}[w]$ into LEX_1 and LEX_2, and see which of the two k-caps activates $\text{PHON}[w]$ when firing back into PHON.

2.2 Results of the word learning experiment

We experiment with different word orders and lexicon sizes, and we input to the system two-word sentences, as described above, *until all three Properties 1-3 above hold.* We are interested in determining how many sentences are needed for this, as a function of the number ℓ of nouns and verbs in the lexicon. Recall that the amount of linguistic input required for language to be learned is of central interest in linguistics, for example, in relation to the "poverty of stimulus" argument [5]. That is, our experiment offers an opportunity to actually quantify what "modest linguistic input" means in the context of language learning. We propose that one useful definition of modest linguistic input is $O(\ell)$ sentences, where ℓ is the size of the lexicon — obviously at least that many sentences are needed.

We find that, indeed, the required number of sentences appears to *grow as a linear function* of the number of words in the lexicon, with a reasonable constant around ten (see Figure 3a). That is, with just around 10 sentences per word, the model (1) converges to stable, self-firing assemblies for each word in the lexicon, (2) correctly classifies nouns vs. verbs by whether an assembly is formed in LEX_1 or LEX_2, and *not* the other, and (3) creating a pathway between PHON, the lexical areas, and the semantic areas that allows for going from one representation to another (e.g. visual stimulus of a dog to PHON["dog"]). We find, perhaps unsurprisingly, that increasing the plasticity parameter speeds up learning, with a natural lower bound in the number of sentences needed to form stable and unstable representations 3b. We also find that the k-cap of a word in the *wrong* lexical area, e.g. from firing PHON["dog"] into LEX_2, becomes increasingly unstable (i.e., less assembly-like) as the number of "complements" that this word occurs with increases; for instance, if "dog" occurs with just 7 different verbs in training, its k-cap into LEX_2 is a set with only 30% stability vs. 100% in LEX_1 (see 3d).

Finally, we run another variant, in which input sentences with intransitive verbs are intermingled with sentences with transitive verbs, such as "the boy pushes the train" containing two nouns. We find that, if the transitive sentences are not more than roughly a half of the input sentences, such a mix of linguistic input also leads to learning of semantics and syntactic roles. In other words, even though the the first experiment is meant for intransitive sentences, the presence of some transitive sentences in the training — which is likely in actual acquisition — does not confound this stage of learning too much.

2.3 Syntax: learning the word order, and learning to generate sentences

After the stage where the meaning and syntactic role (noun vs. verb) of words has been learned, we turn to the key problem of learning *syntax.* Syntax encompasses an entire complex of interrelated linguistic phenomena, and of these we shall focus here on one that is particularly significant: we want our system to learn the language's basic *word order*. To have learned the word order means that the system can *produce novel sentences* when appropriate sensorimotor stimuli are presented — for example, in English, to articulate the sentence "Sister pushes Dad" upon sensory input of the corresponding scene, even if this scene was never seen or described before.

Word orders and grammatical moods. Languages can be broadly classified as belonging to one of six basic categories based on the default ordering of the sentence's subject, object, and verb, known as the *constituent order* or *word order.* For instance, English is an SVO (subject-verb-object) language and so is French, while Japanese is SOV and Tagalog is VSO. Also of interest is the constituent ordering in intransitive sentences (SV vs. VS). The word order is in many ways the most important syntactic characteristic of a language, and is known to determine a large part of the systematic variation in language acquisition [11].

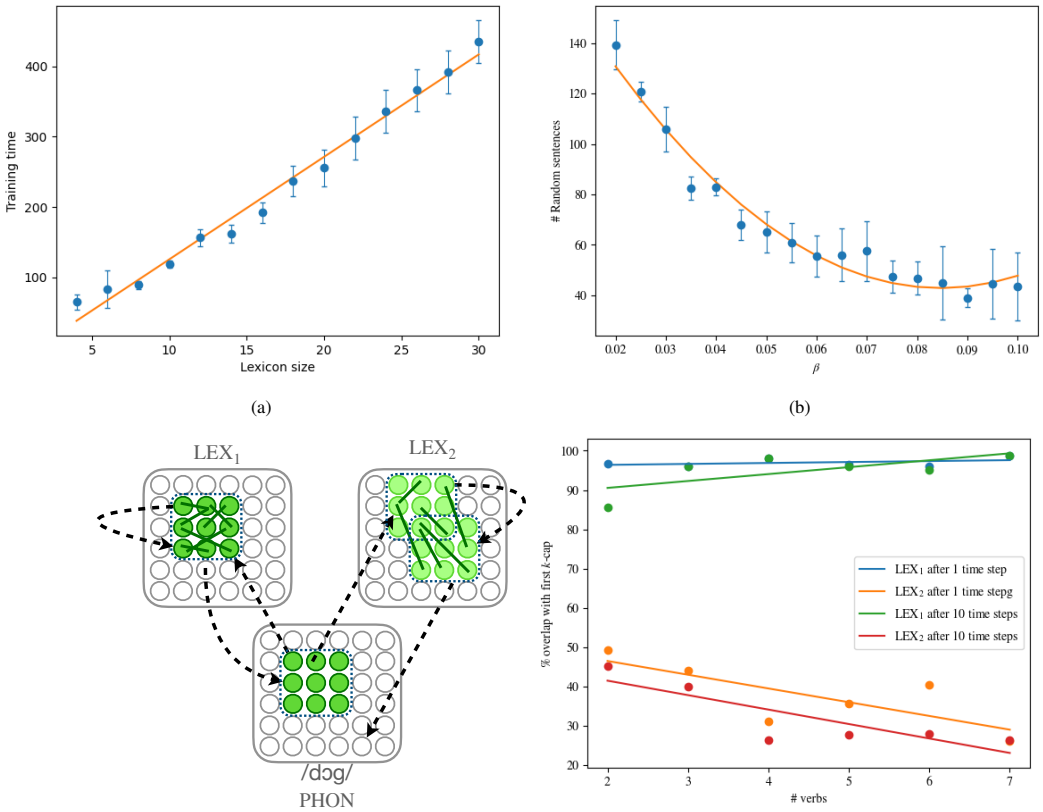

Figure 3: Results of the word learning experiment. In (a) the learning simulation is performed for varying sizes of the lexicon until all of Properties 1-3 hold for the first time for every word, revealing a linear trend (parameters: $n = 10^5, p = 0.05, \beta = 0.06, k_{\text{LEX}_2} = k_{\text{LEX}_1} = 50, k_{\text{CONTEXT}_i} = 20$, and $k = 100$ for other areas, and $\tau = 2$). In (b) the learning experiment is repeated for varying β always for a lexicon of size 4. Learning is accelerated with higher plasticity until a point of saturation. (c) illustrates the formation of a stable assembly (dark green) representing "dog" in LEX_1; in contrast, an unstable and loosely connected set of neurons (pale green) is formed in LEX_2. In (d) the same stability/instability pattern is illustrated numerically, for a varying number of verbs with which "dog" is paired in training sentences. We compare the k-caps obtained by firing PHON[dog], into LEX_1 and LEX_2, and then continuing to fire (both PHON[dog] and the LEX areas recurrently). Stability is measured in terms of the overlap of the k-cap obtained at the first firing of PHON$[dog]$ and that of subsequent firings (one step and ten steps later). The set of firing neurons in LEX_1 is stable, overlapping fully with the first k-cap, while in LEX_2 the overlap is much lower and decreases with more input verbs.

We shall discuss the subject of word orders extensively, but let us note here two apparent linguistic universals [10]: (a) Word orders starting with O are extremely rare; and (b) in attested languages with a rigid or default word order, the intransitive order is almost always the same ordering with O omitted[1] — for example, if the default transitive order is SOV, the intransitive order is SV. As we shall see, these two linguistic facts *can be predicted by the workings of our device:* our device has more difficulty learning the syntax of word orders starting with O, and of languages that violate the second property.

We also note that many languages have varying constituent orders depending on semantic factors such as *mood* — whether the sentence is a statement, question, command, etc. —, syntactic factors (e.g., as part of a dependent clause), or just as a result of emphasis and pragmatics (especially in languages with relatively free word order such as Polish). While we do not set out to handle the entire scope of these complexities, we will show that our model can acquire and output different word orders for

[1]Based off the World Atlas of Language Structures Online [10], out of 1185 languages with default transitive and intransitive orders recorded, we found at most 11 that *may* invert S and V in intransitive order relative to transitive order.

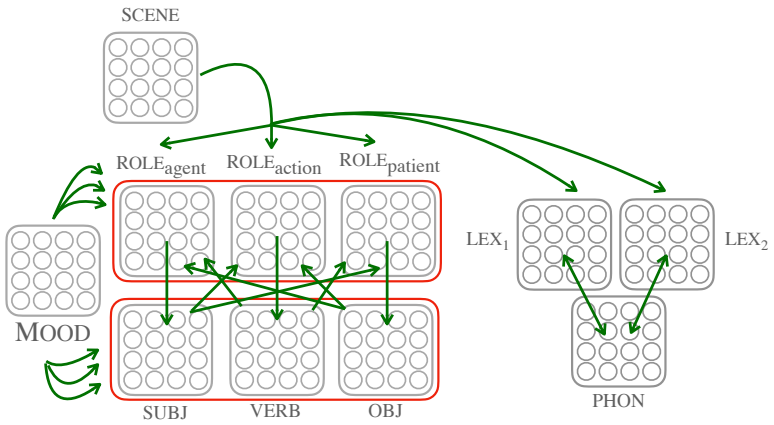

Figure 4: The NEMO system for constituent order learning, as described in 2.3. Red boxes represent two triples of areas under mutual-inhibition. The short arrows out of MOOD depict fibers into each TPJ area, as well as SUBJ, VERB, OBJ. The semantic areas (VISUAL, MOTOR, etc.) are still part of the model, but are not shown for clarity.

different grammatical moods (declarative, interrogative, subjunctive, etc.) via a simple mechanism; we believe that this mechanism can be extended to model syntactic and pragmatic variations of the word order.

Thematic roles. Analyzing a complex scene by assigning thematic roles to the participants of the scene — determining who is doing what to whom, and who is watching — is an important skill for animal survival. A classical study by Norman Geschwind in 1970 [17] showed that Broca's aphasiacs easily select the correct image for a sentence like "the lion chases the deer" and struggle to do so for "the lion chases the tiger," which is consistent with a separate computation of roles in a scene based on visual cues, which helps in the former case but is confounded in the second. The precise seat of the non-linguistic mechanism that assigns roles to perceived agents in a scene is unknown, but a related kind of computation takes place in the left temporoparietal junction (lTPJ), which is believed to be a multimodal hub integrating information from diverse perceptual pathways to "create coherent, spatiotemporal representations of the complex dynamic situations" [3]. The TPJ is often identified as the seat of theory-of-mind in the brain [35]: computing the knowledge states of participants and perspective taking; note that the computation of thematic roles in an observed scene is related to theory-of-mind. A recent metareview of the TPJ [20] concludes that the TPJ is not very involved during conscious reflection about who did what to whom; rather, it appears to carry out a lower-order and automatic processing of perceptual inputs (visual, motor, etc.) in a domain-general way, that is, as the basis for other processes in the brain, e.g. language and social cognition). It was found in [31] that the left TPJ is one of three brain areas that were implicated in the structural analysis of sentences with meaningful words, and not with jabberwocky words. Stephen Pinker hypothesized [33] that the TPJ may be the seat of "mentalese", the nonverbal language of thought. Humans appear to have evolved sophisticated mentalizing abilities through a proliferation in grey matter volume in the temporal cortex, and the TPJ is the most enlarged part of the temporal cortex in humans versus apes [4]. It would seem that the TPJ would be most involved in mental scene "preparation", that is, assigning the roles of agent, patient, observer, etc. to an internal mental scene (perhaps in preparation for speech). The analogous structure in macaques, called SPS, primarily codes attention to other participants, arguably a subtask of role assignment and of the theory of mind [25].

In view of the discussion above, we hypothesize that there is a system in the primate brain, not inherently linguistic and perhaps in communication with the TPJ, which computes thematic roles of scenes. Our NEMO implementation contains three role areas, ROLE$_{agent}$, ROLE$_{action}$, and ROLE$_{patient}$, as well as a fourth area, ROLE$_{scene}$ synaptically connected to all three. We assume that, when a scene is presented, such as a dog chasing a cat, the four areas are prepared as follows: There are assemblies in the three role areas ROLE$_{agent}$, ROLE$_{action}$, and ROLE$_{patient}$ synaptically connected to the assemblies for "dog," "chase," and "cat" in LEX$_1$ and LEX$_2$ (or indirectly thought VISUAL

and MOTOR, respectively), while $\text{ROLE}_{\text{scene}}$ contains an assembly, representing the whole scene, synaptically connected to these three.

The question we shall pursue is this: Assuming that the ROLE areas prepare scene representations as described above (as we presume can be done in all primates), can the constituent order of a human language be learned from input sentences, so as to correctly generate sentences when a new scene is observed or imagined?

The full language system. The full system (Figure 4) has several areas in addition to those of the core word learning system of Figure 2: (a) the four ROLE *areas:* $\text{ROLE}_{\text{agent}}$, $\text{ROLE}_{\text{patient}}$, $\text{ROLE}_{\text{action}}$, and $\text{ROLE}_{\text{scene}}$; (b) three *syntactic* areas, SUBJ, OBJ and VERB, and (c) a MOOD area. There are 18 fibers, shown in Figure 4, which connect the lexical areas to and from ROLE areas, the ROLE areas to the syntactic areas, and MOOD with the ROLE areas and syntactic areas. Finally, the three ROLE areas are in *mutual inhibition;* this is the only use of interarea inhibition in our model.

2.4 The syntax learning experiment

In this section we assume the word learning experiment of Section 2.1 has previously taken place[2]. For learning constituent order, we input grounded sentences, both transitive and intransitive. Grounding now means that, with each input sentence, e.g. "Sister grabs the toy," a scene to the same effect has been input to the sensory system, and the thematic roles (patient, agent, action) of the scene have been computed and are represented as assemblies in the corresponding ROLE areas, with strong synaptic connections to the corresponding assemblies in LEX_1 and LEX_2 (the latter, possibly through the context areas).

For each consecutive word in the sentence, the corresponding assembly in PHON is fired for τ time steps; this allows firing to propagate throughout the lexical areas, ROLE areas, and finally into the syntactic areas. The constituent order is learned through plasticity in the synapses between the syntactic areas and the ROLE role areas. For instance, if the first constituent is the subject, then the firing of the noun's assembly in PHON propagates through LEX_1, to $\text{ROLE}_{\text{agent}}$, and finally into SUBJ. This chain of assemblies continues to fire until the next word, say the verb, is input. Again starting from PHON, the firing of the verb's phonological assembly will propagate through firing into $\text{ROLE}_{\text{action}}$; at this point, neurons from SUBJ (from the previous word) also fire into $\text{ROLE}_{\text{action}}$, and this is how the fact that verb comes after subject is recorded in the synapses between SUBJ and $\text{ROLE}_{\text{action}}$.

The experiment. We fix a language and a mood of the language, together determining one of the six possible word orders, and we input a sequence of random transitive and intransitive sentences with this order. For the experiment to be successful, the proportion of intransitive sentences should not be too high (see the results in Figure 5c); this is to be expected, since an overwhelmingly intransitive input — and consequently a dearth of the patient constituent — is obviously detrimental to the learning of the word order. Every 10 sentences, we probe by running a generation experiment: we test, upon the presentation of a randomly generated novel scene (e.g., a dog eating a cookie, presented as three assemblies in the three ROLE areas), whether the device will generate the appropriate sentence with the correct word order: "the dog eats the cookie" (the details of generation are presented next). Success of the overall learning experiment for this mood and word order is defined as the successful generation of a transitive sentence and another intransitive sentence that are both sampled randomly and withheld during training.

Generation. All previously described functions of the NEMO language system were *reactive,* in the sense that they were initiated by external stimuli: word assemblies firing and sensory data displayed. Generation is different; for example, the human brain has the ability and means to *not* generate the sentence "the dog runs" even though the dog *is* running. Therefore, generation must involve processes, presumably in the prefrontal cortex, implementing the decision to generate. We shall model these processes in the simplest possible way: we assume that there is an assembly, in a new area, which, by firing, inhibits all three ROLE role areas for a single step. We call this assembly the *trigger*.

[2]This is done for the sake of clarity; the two experiments can be intertwined at the cost of more training, as shown in the SI.

We begin with a scene (imagined or seen) presented in the same way as in training: a lexical assembly for each thematic role has been projected into the respective ROLE area, with an assembly in $\text{ROLE}_{\text{scene}}$ synaptically connected to them. Also, the intended mood is determined by an assembly in the MOOD area. These two assemblies, in MOOD and SCENE, keep firing at every step throughout generation. Besides that, the generation algorithm is simply the following.

repeat:

to generate the next word, fire the trigger assembly

until the whole sentence has been generated

Alternatively, generation could be implemented by firing the trigger only once, with PHON causing the trigger to fire through synaptic connections after every generated word.

The simplicity of the algorithm belies quite a bit of complexity. As soon as the ROLE role areas are disinhibited after the single inhibition step, they will receive synaptic input from either (1) mood, (in the case of the first word) or (2) one of the syntactic areas. The role area with the most synaptic input will be selected, and *this will be the next constituent to fire,* an effect achieved through plasticity during the phase of syntax learning. This constituent activates another assembly in the syntactic areas which has highest synaptic connectivity to the next constituent. However, the current constituent will continue to receive the most input from its own recurrent firing, and because the role areas are in mutual inhibition, will continue to fire until the next time all the role areas are inhibited. This will happen as soon as the next word is about to be generated.

2.5 Results of the experiment on word order and generation

We determine how many sentence presentations are needed in order for the model to be able to generate sentences with the correct word order in each mood, for randomly chosen subject, verb, and object, as a function of the size of the lexicon and the number of moods. We withhold one particular transitive sentence and one intransitive sentence from the training set, and test after each training sentence whether the model generates the withheld sentences in the correct order if we set up the corresponding scene in the ROLE areas and generate. The results of our experiments are shown in Figure 5. The device succeeds in learning the constituent order, for several languages and moods with different word orders, after the presentation of a number of sentences that appears to grow linearly with the number of moods, and with the size of the lexicon (see Figure 5b).

The word orders of thousands of languages in the world have been studied extensively, see, for example, the atlas [10], and it is of interest to revisit this knowledge in light of these results. One striking observation from the data already mentioned is that of the six possible word orders, the two starting with object — OVS and OSV — are by far the two rarest, with VOS coming a close third. This fact would be easy to explain if the human brain's system for learning word orders were structured as the one presented here: Our experiments show that these three word orders are more difficult to learn, in the sense that many more input sentences are needed (5c), and a training corpus that is especially rich in transitive verbs is required (something that is not needed for the other word orders). *Ergative languages*, in which, in intransitive sentences, the role of object and subject is reversed, do not present our system with extra difficulties.

Finally, there are many languages (about 17% of the languages in [10]) with no fixed word order, in which the order of the constituents of a sentence is determined by considerations of emphasis and style, as well as languages in which a dominant word order exists but is less strict. We believe that a simple extension of our system can handle such languages through randomization and via the mechanism of additional inputs initiating distinct order chains (as the different MOOD assemblies do in the current model).

2.6 Hierarchy in Language.

So far, we have treated language acquisition as learning how to comprehend and generate sentences, that is, short *sequences* of words. However, it is generally believed that language also has a dual *hierarchical* nature. At the most basic level, each sentence has a *syntax tree* that reveals its syntactic structure (see Figure 6). Furthermore, sentences can be embedded recursively in one another, resulting in larger hierarchies. For example, the utterance *"Melville, whom I love, wrote a book on whaling*

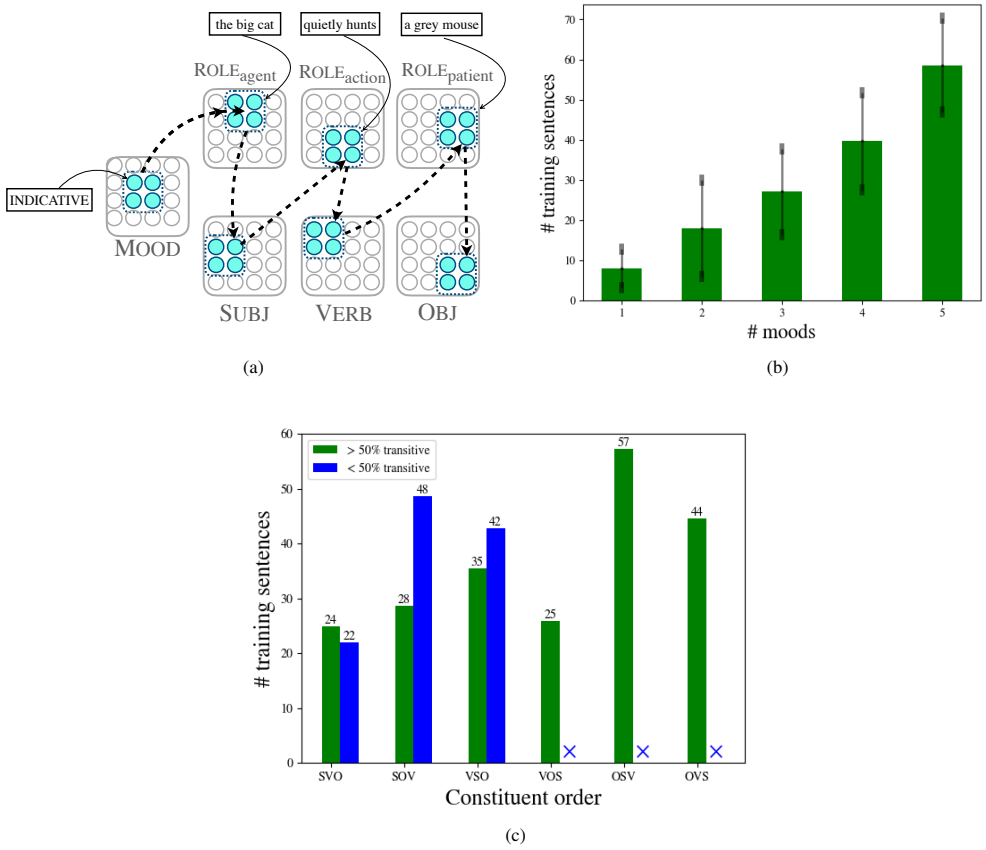

Figure 5: (a) An example of the chain of assemblies that fires in a SVO language: the firing of an assembly in MOOD corresponding to the indicative mood activates the ROLE area with highest input, in this case ROLE$_{agent}$; the assembly in ROLE$_{agent}$ fires, activating an assembly in SUBJ. When the trigger assembly is fired, inhibiting all three ROLE areas, the next one to fire is the one with greatest input from SUBJ: ROLE$_{action}$, and so on. In the labels of this example, we imagine an expanded system where the assemblies in each ROLE area can represent whole *phrases,* to suggest how our algorithm for constituent order learning and production is not confined to 3-word sentences. (b) Our system can learn languages with multiple word-orders corresponding to distinct grammatical moods (each mood corresponding to a distinct assembly in the MOOD area). Intuitively, each mood corresponds to a distinct *chain* of assemblies between ROLE and SUBJ, VERB, OBJ. We average over 10 runs per number-of-moods, revealing a roughly linear trend, with somewhat steeper increases in the difficulty to learn the 4th and the 5th mood. (c) We compare the number of sentences needed to learn different constituent orders. The algorithm succeeds for any order if more transitive sentences are presented than intransitive sentences, but we find that significantly more sentences are needed to learn OSV and OVS (averaging over 20 runs per order, with 70% transitive input). If the proportion of transitive sentences is less than half, it is easy to show via rigorous proof that the orders VOS, OSV, or OVS *cannot* be learned by this system even with unbounded input, and so either an altogether different mechanism, or predominance of transitive input sentences must account for the acquisition of syntax in such languages — note that these are in fact the three rarest constituent orders among world languages.

that became a classic" is the nested combination of three individual sentences (see Figure 6a). The relative importance of the hierarchical and sequential aspects of language is the subject of a well-known discord in linguistics; see, for example, [9], as well as the responses to that paper, for a recent recurrence.

Can our implementation of language acquisition shed light on the role of hierarchies in language? To quote noted neurolinguist Angela Friederici [16], *"It would really be crucial to show that there is a limit for ordered sequences and that beyond that limit a (hierarchical) nesting is necessary."*. Interestingly, in recent work on the implementation in NEMO of planning [12] and sequence memorization [6] it has been observed that, precisely as Friederici has suggested, it is difficult to maintain sequences of assemblies longer than a limit. The limit varies with the parameters of the NEMO model, and ranges between 20 and 40. This difficulty suggests that a full language system implemented in NEMO must also deal with the hierarchical aspects of language.

The language acquisition system presented so far can be extended in a very natural way to support the creation of hierarchical structures. Much of the system in Figure 4 — that is, the areas for phonology, verbs, nouns, moods, and roles, omitting the semantic areas — models parts of the human language system that are thought to be located in the left medial temporal lobe (lMTL) of the brain, in what is broadly referred to as *Wernicke's area*. Now we can postulate that the three syntactic areas (SUBJ, VERB, OBJ) are located in *Broca's area* in the left frontal gyrus; recall that parts of Broca's area are thought to be involved in syntactic processing and generation [15], and the two areas of Broca and Wernicke are connected by the arcuate fasciculus (AF) axon bundle. That is, we postulate that the nine fibers connecting the role areas and the syntactic areas in Figure 4, implementing word order, are parts of the AF. In addition to the three areas SUBJ, VERB, and OBJ, we can assume two more areas named SENT and VP, also meant to be located in Broca's area, standing for *sentence* and *verb phrase,* respectively. These five areas are connected through fibers as follows: VERB and OBJ are both connected to VP, while SUBJ and VP are both connected to SENT (see Figure 6b). This structure mirrors the skeleton of the basic syntax tree of a transitive sentence, and the same structure can also work for intransitive sentences. The effect of this new arrangement is that, during both comprehension and production of a sentence, the basic syntax tree of the sentence is created through synaptic connections between the assemblies in these areas. As for embedded structure, like the relative clauses of the sentence in Figure 6a, previous work on the theory of parsing in NEMO showed that such a structure can be handled with an additional area (called DEP in previous work) whose assemblies link a word to its dependent clause [27]. However, parsing and generating such hierarchically structured sentences presents novel challenges, which need further work.

3 Discussion

We presented what we believe is the first biologically plausible system that can learn the rudiments of natural language — lexicon as well as basic syntax and semantics — through the presentation of a modest amount of grounded sentences in any language. Of course, many facets of language are not covered by our system: The learning of speech sounds and phonology has been bypassed, while adjectives, adverbs, and functional categories are not considered. We believe that many of these missing aspects can be accommodated by simple extensions of the present system. *Abstract words* — words that are not directly grounded in sensorimotor experience, such as "bravery," "available," and "negotiate" — are an important next step that will require, among others, the development of a representation scheme for such words and the introduction of new semantic brain areas. There are many current cognitive linguistic theories on how abstract words are dealt with in the human brain, see, for example, [2], which can guide progress on this front. Other sophisticated aspects of linguistic competence, such as resolution of ambiguity, grammaticality judgment, metaphor, and next-word prediction, to mention a few, are very interesting further challenges, and we are confident that they can be addressed in our framework, even though some of them may require certain simple and biologically plausible extensions of the NEMO model (recall the discussion of moods in the previous section).

Further research goals would be learning the social uses of language, as well as its interaction with aspects of social cognition such as agency and the theory of mind, which presuppose new kinds of representations and mechanisms. In addition, it would be very intriguing to experiment with interactions between NEMO artifacts and large language models (LLMs) as a means to help the

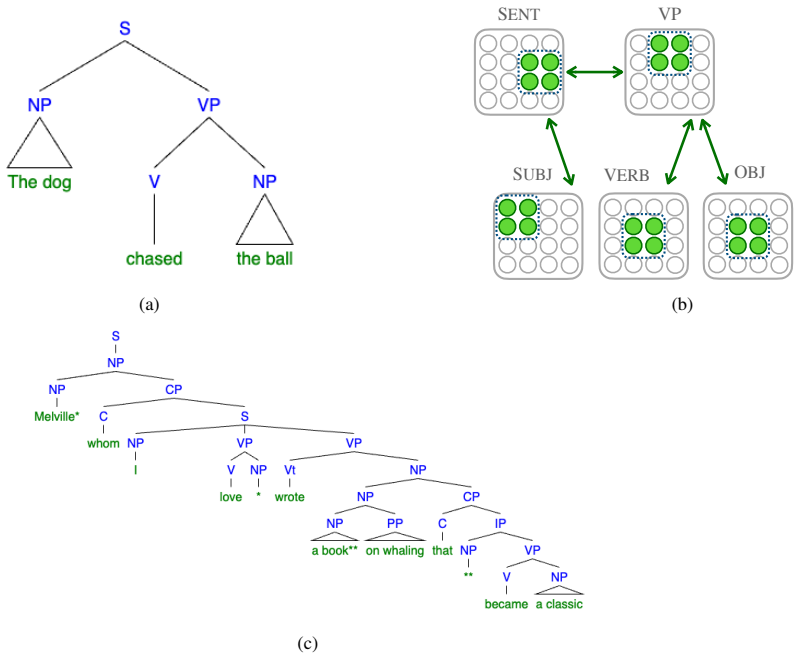

Figure 6: (a) The syntax tree of a simple sentence, with subject noun phrase, and verb phrase composed of a verb and object noun phrase. (b) An extension of the language acquisition system that can create hierarchical structure (namely, representations of the verb phrase (VP) and of the whole sentence) through the addition of two additional areas (VP and SENT, representing subareas of Broca's area. Fibers are shown as green arrows. Firing of assemblies in VERB and OBJ into VP activates an assembly representing the verb phrase; similarly, firing from VP and SUBJ into SENT results in a firing assembly representing the whole sentence (or the "root" of the syntactic tree at S in subfigure (a)). (c) The syntax tree of a sentence, showing a hierarchy of embedded structure.

former grow in these more demanding directions. Beyond language, *reasoning* in its various genres — logical, heuristic, categorical, probabilistic, etc. – is an important next challenge for NEMO artifacts.

Our system can be extended to handle *multilinguality:* A new area can be added containing an assembly for each language known to the speaker, very much like the representations of grammatical moods in the MOOD area. This assembly would be active while learning, or using, the corresponding language. Initial experimentation with this idea is encouraging, but more work is needed.

We believe that our language acquisition system is biologically plausible not only because it is strictly based on the NEMO framework, but also because its *tabula rasa* brain architecture — the areas and the fibers connecting them — can be plausibly delivered by mammalian developmental processes. One possible exception are the assemblies in the MOOD area, each representing a different mood of the language. It can be argued that such a mood assembly can be created during learning on the basis of various sensory and mental-state cues associated with the mood (e.g., the tension in the discourse and the special tone of voice when a question is asked). However, more work is needed to fully justify this point.

Notwithstanding the important and challenging research directions that lie ahead and the limitations mentioned above, the artificial system for basic language acquisition based on NEMO we present here is the first of its kind, and, besides its value as a proof of concept, may be of use to neurolinguistic research. Its design decisions are intended to largely reflect current neurolinguistic thinking, and hence the resulting artifact is a concrete hypothesis that can possibly help guide further research on the human language system through the generation of testable predictions and the deployment of more advanced versions.

References

[1] Richard Axel. 2018. Q & a. *Neuron*, 99:1110–1112.

[2] Anna M. Borghi, Ferdinand Binkofski, Cristiano Castelfranchi, Felice Cimatti, Claudia Scorolli, and Luca Tummolini. 2017. The challenge of abstract concepts. *Psychological bulletin*, 143 3:263–292.

[3] Christine Boylan, John C. Trueswell, and Sharon L. Thompson-Schill. 2015. Compositionality and the angular gyrus: A multi-voxel similarity analysis of the semantic composition of nouns and verbs. *Neuropsychologia*, 78:130–141.

[4] Marius Braunsdorf, Guilherme Blazquez Freches, Lea Roumazeilles, Nicole Eichert, Matthias Schurz, Sebo Uithol, Katherine L. Bryant, and Rogier B. Mars. 2021. Does the temporal cortex make us human? a review of structural and functional diversity of the primate temporal lobe. *Neuroscience & Biobehavioral Reviews*, 131:400–410.

[5] Noam Chomsky. 1986. *Knowledge of Language: Its Nature, Origin, and Use*. Prager.

[6] Max Dabagia, Christos Papadimitriou, and Santosh Vempala. 2024. Computation with sequences of assemblies in a model of the brain. *Neural Computation*, 37:193–233.

[7] Max Dabagia, Santosh S Vempala, and Christos Papadimitriou. 2022. Assemblies of neurons learn to classify well-separated distributions. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 3685–3717. PMLR.

[8] Stanislas Dehaene, Fosca Al Roumi, Yair Lakretz, Samuel Planton, and Mathias Sable-Meyer. 2022. Symbols and mental programs: a hypothesis about human singularity. *Trends Cogn Sci.*, pages 751–766.

[9] Nai Ding, Lucia Melloni, Hang Zhang, Xing Tian, and David Poeppel. 2016. Cortical tracking of hierarchical linguistic structures in connected speech. *Nature neuroscience*, 19(1):158.

[10] Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.4)*. Zenodo.

[11] Cristina D. Dye, Yarden Kedar, and Barbara Lust. 2018. From lexical to functional categories: New foundations for the study of language development. *First Language*, 39:32 – 9.

[12] Francesco d'Amore, Daniel Mitropolsky, Pierluigi Crescenzi, Emanuele Natale, and Christos H. Papadimitriou. 2022. Planning with biological neurons and synapses. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(1):21–28.

[13] Leonardo Fernandino and Marco Iacoboni. 2010. Are cortical motor maps based on body parts or coordinated actions? implications for embodied semantics. *Brain and Language*, 112(1):44–53. Mirror Neurons: Prospects and Problems for the Neurobiology of Language.

[14] Angela D. Friederici. 2009. Pathways to language: fiber tracts in the human brain. *Trends in Cognitive Sciences*, 13(4):175–181.

[15] Angela D Friederici. 2017. *Language in our brain: The origins of a uniquely human capacity*. MIT Press.

[16] Angela D. Friederici and Noam Chomsky. 2017. *Language in Our Brain: The Origins of a Uniquely Human Capacity*. The MIT Press.

[17] Norman Geschwind. 1970. The organization of language and the brain. *Science*, 170(3961):940–944.

[18] Albert Gidon, Timothy Adam Zolnik, Pawel Fidzinski, Felix Bolduan, Athanasia Papoutsi, Panayiota Poirazi, Martin Holtkamp, Imre Vida, and Matthew Evan Larkum. 2020. Dendritic action potentials and computation in human layer 2/3 cortical neurons. *Science*, 367(6473):83–87.

[19] Gregory Hickok, Bradley Buchsbaum, Colin Humphries, and Tugan Muftuler. 2003. Auditory–motor interaction revealed by fmri: Speech, music, and working memory in area spt. *J. Cognitive Neuroscience*, 15(5):673–682.

[20] Kajsa M. Igelström and Michael S.A. Graziano. 2017. The inferior parietal lobule and temporoparietal junction: A network perspective. *Neuropsychologia*, 105:70–83. Special Issue: Concepts, Actions and Objects: Functional and Neural Perspectives.

[21] Eric R. Kandel, James H. Schwartz, and Thomas M. Jessell, editors. 1991. *Principles of Neural Science*, fifth edition. Elsevier, New York.

[22] Kohitij Kar and James J. DiCarlo. 2024. The quest for an integrated set of neural mechanisms underlying object recognition in primates. *Annual Review of Vision Science*, 10:91–121.

[23] David Kemmerer and Javier Gonzalez-Castillo. 2010. The two-level theory of verb meaning: An approach to integrating the semantics of action with the mirror neuron system. *Brain and Language*, 112:54–76.

[24] David Kemmerer, David Rudrauf, Kenneth Manzel, and Daniel Tranel. 2012. Behavioral patterns and lesion sites associated with impaired processing of lexical and conceptual knowledge of actions. *Cortex; a journal devoted to the study of the nervous system and behavior*, 48:826–48.

[25] Rogier B. Mars, Jérôme Sallet, Franz-Xaver Neubert, and Matthew F. S. Rushworth. 2013. Connectivity profiles reveal the relationship between brain areas for social cognition in human and monkey temporoparietal cortex. *Proceedings of the National Academy of Sciences*, 110(26):10806–10811.

[26] Daniel Mitropolsky, Michael J. Collins, and Christos H. Papadimitriou. 2021. A Biologically Plausible Parser. In *Transactions of the Association for Computational Linguistics*. ArXiv: 2108.02189.

[27] Daniel Mitropolsky, Adiba Ejaz, Mirah Shi, Christos Papadimitriou, and Mihalis Yannakakis. 2022. Center-embedding and constituency in the brain and a new characterization of context-free languages. In *Proceedings of the 3rd Natural Logic Meets Machine Learning Workshop (NALOMA III)*, pages 26–37, Galway, Ireland. Association for Computational Linguistics.

[28] Simone Mätzig, Judit Druks, Jackie Masterson, and Gabriella Vigliocco. 2009. Noun and verb differences in picture naming: Past studies and new evidence. *Cortex*, 45(6):738–758.

[29] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2024. A comprehensive overview of large language models.

[30] Kayoko Okada and Gregory Hickok. 2006. Left posterior auditory-related cortices participate both in speech perception and speech production: Neural overlap revealed by fmri. *Brain and Language*, 98(1):112–117.

[31] Christophe Pallier, Anne-Dominique Devauchelle, and Stanislas Dehaene. 2011. Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, 108(6):2522–2527.

[32] Christos H. Papadimitriou, Santosh S. Vempala, Daniel Mitropolsky, Michael Collins, and Wolfgang Maass. 2020. Brain computation by assemblies of neurons. *Proceedings of the National Academy of Sciences*, 117(25):14464–14472.

[33] S. Pinker. 2010. *The Language Instinct: How The Mind Creates Language*. HarperCollins.

[34] Sara Popham, Alexander Huth, Natalia Bilenko, Fatma Deniz, James Gao, Anwar Nunez-Elizalde, and Jack Gallant. 2021. Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature Neuroscience*, 24:1628–1636.

[35] R Saxe and N Kanwisher. 2003. People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind". *NeuroImage*, 19(4):1835–1842.

[36] Mark S. Seidenberg and Laura A. Petitto. 1979. Signing behavior in apes: A critical review. *Cognition*, 7(2):177–215.

[37] Rosario Tomasello, Max Garagnani, Thomas Wennekers, and F Pulvermueller. 2018. A neurobiologically constrained cortex model of semantic grounding with spiking neurons and brain-like connectivity. *Frontiers in Computational Neuroscience*, 12(0).

[38] Gabriella Vigliocco, David P. Vinson, Judit Druks, Horacio Barber, and Stefano F. Cappa. 2011. Nouns and verbs in the brain: A review of behavioural, electrophysiological, neuropsychological and imaging studies. *Neuroscience & Biobehavioral Reviews*, 35(3):407–426.