Encoding Words into a Potts Attractor Network

Sahar Pirmoradian, Alessandro Treves
SISSA, Cognitive Neuroscience Sector, Trieste, Italy

To understand the brain mechanisms underlying language phenomena, and sentence construction in particular, a number of approaches have been followed that are based on artificial neural networks, where words are encoded as distributed patterns of activity. Still, issues like the distinct encoding of semantic vs syntactic features, word binding, and the learning processes through which words come to be encoded that way, have remained tough challenges. We explore a novel approach to address these challenges, which focuses first on encoding words of an artificial language of intermediate complexity (BLISS) into a Potts attractor net. Such a network has the capability to spontaneously latch between attractor states, offering a simplified cortical model of sentence production. The network stores the BLISS vocabulary, and hopefully its grammar, in its semantic and syntactic subnetworks. Function and content words are encoded differently on the two subnetworks, as suggested by neuropsychological findings. We propose that a next step might describe the self-organization of a comparable representation of words through a model of a learning process.

Keywords: Word Representation, Artificial Language, Potts Attractor Network

1. Introduction

To understand the brain mechanisms underlying the phenomenon of language, specifically sentence construction, many studies have been done to implement an artificial neural network that encodes words and constructs sentences (see e.g. [1–4]). These attempts differ on how the sentence constituents (parts) are represented—either locally [1,3], or in a distributed fashion [5,6]—and on how these constituents are bound together—through either temporal synchrony [7], active circuits [3], or algebraic operations [8].

The local representation of each sentence constituent (either a word, a phrase, or even a proposition) results in an exponential growth in the number of units needed for structure representation [1]; this challenge was addressed in [3] by designing dynamic circuits between word assemblies, yet with a highly complex and meticulously (unrealistic) organized connections. In a fully distributed representation of words as vectors [5,6], words are bound

(and *merged*) together by an algebraic operation—e.g. tensor product [8] or circular convolution [6]. Some steps have been attempted towards the neural implementation of such operations [4]. Another distributed approach was towards implementing a simple recurrent neural network that predicts the next word in a sentence [9]. Apart from the limited language size that the network could deal with [10], this system lacked an explicit representation of syntactic constituents, and it is shown it leads to a lack of grammatical knowledge in the network [11,3].

However, despite all these attempts, there remains the lack of a neural model that addresses the challenges of language size, semantic and syntactic distinction, word binding, and word implementation in a neurally plausible manner. We are exploring a novel approach to address these challenges, that involves encoding words of an artificial language of intermediate complexity into a neural network, as a simplified cortical model of sentence production, which stores the vocabulary and the grammar of the artificial language in a neurally plausible manner on two components: one semantic and one syntactic.

2. BLISS: The Training Language

As the training language of the network, we have constructed *BLISS* [12], for Basic Language Incorporating Syntax and Semantics. BLISS is a scaled-down synthetic language of intermediate complexity, which mimics natural languages by having a vocabulary, syntax, and semantics. Importantly, the degree of complexity of the language is designed having the size limitations of synthetic agents in mind, so as to allow for the use of equivalent corpora with human subjects and with computers, while aiming for reasonable linguistic plausibility.

BLISS is generated by a context-free grammar of limited complexity with about 40 production rules, with probabilities that were drawn from the *Wall Street Journal* (WSJ) corpus. It contains function words, inflectional suffixes, and some embedding structure. These grammatical features were introduced to enable experiments to investigate the ability for abstract pattern acquisition [13,14], the special role of function words [15], the role of suffixes [16], and especially hierarchical structures [17,18] in humans.

The BLISS vocabulary contains about 150 words, which belong to different lexical categories such as noun, verb, adjective, etc., and which were selected from the *Shakespeare* corpus. There are several studies investigating category learning in humans [19], and BLISS is intended to facilitate e.g. the analysis of the representation of distinct lexical categories.

Semantics is defined in BLISS as the statistical dependence of each word on other words in the same sentence, as determined solely by imposing constraints on word choice during sentence generation. We have applied different methods of weighing the preceding words so to determine which words come next. This should allow using BLISS to study at least rudimentary aspects of the emergence of semantic categories.

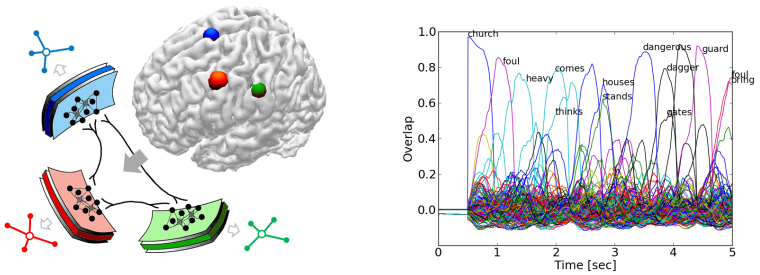
3. Potts Attractor Network: a Simplified Model of the Cortex

We have attempted to implement a neural network which mimics the neural mechanisms underlying sentence production. We use a *Potts associative memory network*, a generalization of an auto-associative memory network, an *attractor network* [20,21].

An attractor network is a collection of binary units that stores a concept—a pattern—in a distributed fashion, remembers a concept by completing a portion of it given as a cue, and uses a Hebbian learning rule to store a concept as an attractor at a minimum of the (free) energy of the network.

In the Potts associative memory network—the network of our interest—the units are not binary; instead, each can be activated in S different states. The Potts network has been proposed as a simplified model of macroscopic cortical dynamics [22,23], perhaps appropriate for modelling the language faculty and other high-level cognitive functions (Fig. 1(a)). The Potts network is a simplified two-level, local and global associative memory network [24,25], where a local network represents a patch of cortex, which locally stores features, and the global network associates those features to store concepts (as proposed by [26]). In the Potts network, the local associative memory networks are not described explicitly; instead, each is encapsulated each as a *Potts* unit. Thus, a Potts unit hypothetically models a patch of cortex, and the internal neuronal dynamics of the patch is not described by the model, rather it is subsumed into an effective description in terms of graded Potts units with adaptation effects. A collection of Potts units, connected through long-range synaptic connections, compose the global associative memory, which stores the concepts.

Apart from the simplification the Potts network offers, the choice of this network for sentence production has been mainly motivated by its "latching" dynamics [27]. Latching is an ability to jump spontaneously and in some conditions indefinitely from an attractor state to the next, in a process that mimics spontaneous language production. This behaviour is illustrated in



(a) Conceptual derivation of the Potts network

(b) An example of latching dynamics in the Potts network

Fig. 1: (a) Conceptual derivation of the Potts network, which models the cortex as a simplified two-level, local and global associative memory network. The local dynamics of a cortical patch (sheets) are reduced to a Potts unit with several states (here, 4). The global associative memory is the collection of the Potts units, which are connected through long range connections (black connections between cortical patches). (b) An example of latching dynamics in the Potts network. The x axis is time and the y axis is the correlation (overlap) of the network state with specific stored patterns. Each memory pattern represents a word. In this simulation the memory patterns are randomly correlated.

Fig. 1(b), which shows the overlap between the actual network activation and the activation pattern that characterises the stored patterns as a function of time. Initially, an externally cued attractor leads to retrieval—a full retrieval corresponds to an overlap of one. However, the activation of the network does not remain in the retrieved pattern. Instead, it jumps or latches from attractor to attractor, driven by adaptation effects. Jumps between attractors are facilitated by an overlap between the current and the subsequent memory pattern.

4. Implementation of Word Representation in the Potts Network

After constructing the training language, BLISS, and implementing the Potts network, we need to represent the BLISS words into the network.

We represented the BLISS words in a distributed fashion on 900 Potts units, among which, 541 units express the semantic content (comprising the *semantic sub-network*), and the rest, 359 units, represent the syntactic characteristics of a word. The distinction between the semantic and syntac-

tic characteristics of a word has been loosely inspired by neuropsychological studies ([28,29]). We have also made a distinction between the representation of function words (e.g. prepositions, determiners, and auxiliary words) and content words (e.g. nouns, verbs, and adjectives), as suggested by several neuropsychological findings [30,31]. While the sparsity (the fraction of active Potts units) was kept the same for all the words ($a = 0.25$) on all 900 units, it is not equally distributed between semantic and syntactic sub-networks: semantic units are less active in the case of the function words (45 active units out of 541 units) compared to the content words (135 out of 541), whereas syntactic units are more active for the function words (180 out of 359) than for the content words (90 out of 359).

To have a word-generating algorithm that reflects the variable degree of correlation between words, we used an algorithm comprised of two steps [23]: (1) we establish a number of vectors called *factors* or features. Each factor influences the activation of some units in a word by "suggesting" a particular state to that unit. A word can be called a child, as it is generated by several factors as parents. (2) The competition among these factors through their *suggestion weights* determines the activation state of each unit in a word. In each unit, the state with the strongest suggestion is the winner. In order to maintain the desired level of sparsity, we picked the units with stronger suggestions in their selected states, and inactivated the remaining units by setting them to the null state.

To determine suggestion weights of a factor for its child, we used, whenever possible, the co-occurrence of the factor and its child in the BLISS corpus generated by *Subject-Verb* model. As we generate each word category in the next sections, we will specify our choice for the suggestion weights.

The algorithm includes a noise term to avoid generating words with very high correlation. We produced a number of additional factors, called *hidden* factors, whose suggestion weights were randomly selected from the distribution of the weights of the visible or main factors.

The proposed word-generating algorithm can be argued to be consistent with the findings of recent fMRI computational studies, which attempted to predict the neural signature of words by considering some other words as features—the factors in our algorithm. For instance, in [32], the fMRI neural representations of some nouns were predicted by proposing a linear model that considered 25 verbs as features. In this study, features compete through weights that correspond to the co-occurrence of the feature and the main noun in a natural language. To test the ability of the model for predicting

words by having a more diverse range of features, they considered 1000 frequent words instead of 25 as the features; the model again succeeded in predicting the fMRI BOLD response of the nouns, though with lower accuracy.

Generating words using the above algorithm, we quantified the correlation between two words as the number of active units that are at the same state (Nas) in both patterns. We use the notation of $< Nas >$ to measure the average correlation across all words of two word categories—either the same or different categories.

If two patterns μ and ν are randomly correlated, we expect the correlation measure to be $Na^{\mu}\frac{a^{\nu}}{S}$, where a^{μ} is the sparsity of pattern μ and S the number of active states that a unit of this pattern may occupy. If we store randomly correlated patterns in the semantic sub-network ($N_{sem} = 541$), this measure reads: $Nas_{sem} \simeq 4.8$ between two words with $a = 0.25$, and $Nas_{sem} \simeq 1.5$ between two words with $a = 0.25$ and $a = 0.08$; and, $<Nas_{sem}> \simeq 4.1$, averaged over all words, if there are 134 words with $a = 0.25$ and 15 words with $a = 0.08$. On the other hand, with randomly correlated patterns in the syntactic sub-network ($N_{syn} = 359$), this measure reads: $Nas_{syn} \simeq 3.2$ between two words with $a = 0.25$, and $Nas_{syn} \simeq 6.4$ between two words with $a = 0.25$ and $a = 0.50$; and $<Nas_{syn}> \simeq 3.3$, averaged over all words, if there are 134 words with $a = 0.25$ and 15 words with $a = 0.50$.

4.1. *Semantic Representation*

To generate the semantic units of words, first nouns were generated using some *feature norms* as factors—feature norms include a list of features for a concept (e.g. *is-animal* and *a-mammal* are the features of *dog*). We then generated adjectives and verbs using nouns as factors. Finally, nouns, adjectives, and verbs served as factors for the generation of proper nouns and function words.

For the representation of nouns we used the feature norms in the McRae database [33]. The database was collected from an experimental study in which 541 nouns, including 18 BLISS nouns, were associated by human participants with a set of feature norms. In total, for all 541 nouns, 2500 features (e.g. *is-made-of-metal*, *is-animal*, *a-mammal*) were used; out of 2500 features, 190 features were associated to 18 BLISS nouns. We represented these features as vectors of 541 elements with $a = 0.25$.

To represent these 190 features as vectors with 541 elements we followed several steps: (1) we sorted the features, $f_1...f_i...f_{190}$, in descending order,

by the number of concepts (or nouns) that are associated per feature, ω_{f_i}, in the database (e.g. $\omega_{an-animal} = 90$); (2) in an orderly fashion, we picked a feature in the list, randomly selected some units of its 541-element vector, then assigned their states by considering the previous features in the list as their factors. The number of randomly selected units was $3 * (33 + \omega_{f_i})$, because $<\omega_{f_i}> = 12$ and we needed to maintain the average sparsity around $a = 0.25$ (135 active units). As the first feature in the list did not have any preceding feature, we randomly assigned the states of its units. The suggestion weight was the co-occurrence frequency of features in the database; hence, the features that are more often associated with the same nouns will have higher correlation.

After the representation of the features of the McRae database, we used these 190 features as factors for the generation of the nouns. For a given noun, the features that are associated with that noun in the McRae database suggest the activation state of the units, with the weight of $\frac{1}{3*(33+\omega_{f_i})}$, to strengthen the uniqueness of the features. Hence the features that are more distinct in the database (e.g. *a-baby-cow* with $\omega = 1$) because of their smaller ω give more distinctive, stronger suggestions to a noun than popular features (e.g. *an-animal* with $\omega = 90$). The features that suggest the states of a noun and are associated with that noun in the database are likely to suggest other nouns that belong to the same semantic category; thus we expect higher correlations between words of the same semantic category.

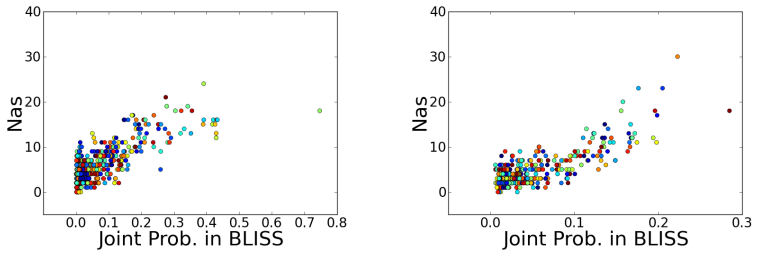
After generating the semantic units of the nouns, we produced the semantic representation of the 37 verbs and the 18 adjectives of BLISS by using the nouns as factors. The suggestion weight of a noun for a verb or an adjective is determined by the co-occurrence probability of the noun and the corresponding word (either verb or adjective) in the BLISS corpus; hence the representation of a verb or an adjective tends to be more correlated with the nouns that appear more frequently with it in the corpus. For the generation of verbs and adjectives we added about 400 hidden factors in addition to their main factors, to avoid high correlations between these words. High correlations would have interfered with the dynamics of the semantic network.

After generating the semantic representation of nouns, verbs, and adjectives, we used these content words as factors—together with 400 hidden factors—to generate 6 proper nouns and then 15 function words.

As for the singular and plural form of words, we assumed that the meaning (the semantic part) should be the same for both numeral forms,

and the only distinction should be in syntactic representation. Therefore, the plural and singular forms of nouns and verbs (e.g. *dog* and *dogs*, or *kill* and *kills*) are stored as identical in the semantic sub-network.

In the semantic representation, a generating factor influences the representation of a word by a weight that is proportional to the joint probability of the factor and the word. We have thus compared the correlation of some factors with the generated word in the Potts network and in the training BLISS corpus, generated by the Subject-Verb model (Fig. 2). This correlation was measured as $<Nas>$ in the Potts network, and as *joint probability* in the BLISS corpus (the joint probabilities between a word and its factors were normalized to 1). Although in the generation of the words, a very high noise level—about 400 hidden factors—was used to decrease the correlation between words, Fig. 2 demonstrates that a highly frequent word pair in the BLISS corpus still has a high correlation (Nas) in the semantic sub-network. These high Nas correlations indicate a deviation from the regime of randomly correlated patterns ($<Nas> \simeq 4.8$, between content words).



(a) Verbs and Nouns (b) Adjectives and Nouns

Fig. 2: Comparison of the correlation between words and their factors in their semantic representation ($<Nas>$ on y-axis) versus their joint probabilities in the BLISS corpus produced by the Subject-Verb model (x-axis) (the joint probabilities between a word and its factors have been normalized to 1). Each dot indicates a pair of a word and its generating factor. (a) The correlation between verbs and nouns (the generating factors of verbs) in their semantic representation vs. the joint probability between the verbs and the nouns in the BLISS corpus (e.g. *kill sword*); Likewise, for (b) adjectives and nouns (adjectives' generating factors) (e.g. *bloody sword*).

4.2. *Syntactic Representation*

For the syntactic representation of words, we first generated function words using a limited set of somewhat arbitrarily designed syntactic features. Using function words as factors together with those syntactic features, we generated the syntactic representation of nouns, verbs, adjectives, and proper nouns.

As factors for generating the function words, we arbitrarily designed 19 syntactic features: 7 lexical categories (noun, verb, adj, conjunction, preposition, pronoun, adverb); 2 numbers (singular, plural); 1 negation; 3 determiners (indefinite, definite, properNoun); 2 locations (close, far); and 4 directions (from, towards, samePlace, above).

We represented the above syntactic features as vectors of 359 elements with $a = 0.25$ (90 active units), while keeping the representation of features within each of the above categories orthogonal to each other. For instance, for the first item, *lexical categories*: (1) we generated the representation of *lxc/noun*, by randomly selecting 90 units and arbitrarily assigning their activation states; (2) for *lxc/verb*, we activated the same units as in the *lxc/noun* but assigning different states; (3) for the rest of the members (i.e. *lxc/adj*, ...), we used the same procedure as in (2) while keeping all these features completely orthogonal. We took the same steps (1)–(3) for other categories listed above while keeping the features within a category uncorrelated.

Since these syntactic features will be used as the factors for all the words, we arbitrarily set their suggestion weights for the generation of different word categories, either function words or content words (see Table 1 for some examples).

We used the above 19 syntactic features, together with 20 hidden factors, as the factors for the syntactic representation of 15 function words. Using the function words and the syntactic features, together with 20 hidden factors, we generated the syntactic representation of 36 nouns (singular and plural), 74 verbs (singular and plural), 18 adjectives, and 6 proper nouns (singular and plural). The suggestion weights of the function words for the generation of a content word are determined by the joint probability of the two corresponding words in the BLISS corpus. Thus, if a content word has a higher co-occurrence with a function word in the corpus, the representations of these two words tend to be more correlated.

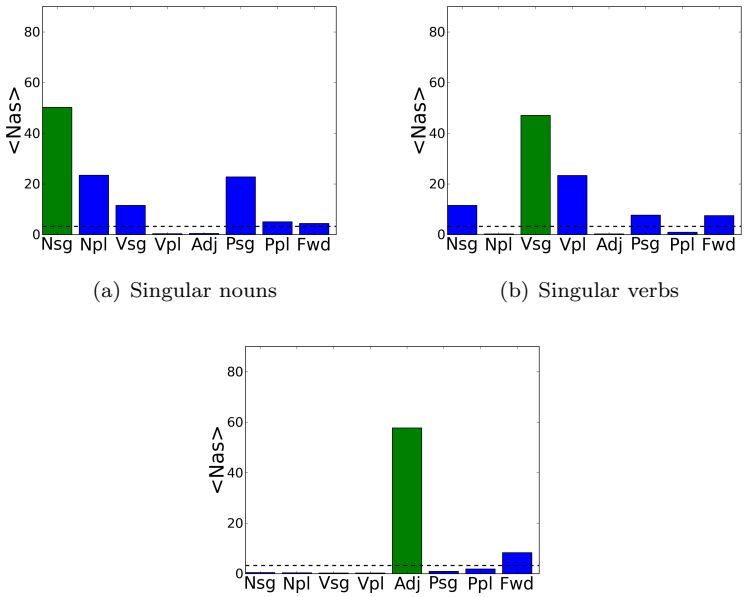
Generating the syntactic representation of all the words, we measured their correlations within and across different syntactic categories (singular and plural nouns, singular and plural verbs, adjectives, singular and plural

proper nouns, and function words), as shown in Fig. 3. As expected, the correlations between relevant syntactic categories highly deviate from the regime of randomly correlated patterns in the syntactic sub-network; for randomly correlated patterns, $<Nas> \simeq 3.2$, between content words, and $<Nas> \simeq 6.4$, between content words and function words. As shown in Fig. 3(a), singular nouns (*Nsg*) have higher correlations with other noun categories (i.e. plural nouns and proper nouns) and also with other singular words (i.e. singular verbs), than with plural verbs or with adjectives. Though function words (*Fwd*) participate as factors in the generation of all the content words, their correlations with other categories are relatively small, even within the function words themselves, because of their high sparsity ($a_{syn}^{fwd} = 0.50$) compared to other words and to their syntactic features.

Using auto- and hetro-associative learning rules [34], we stored the implemented words into the network. Fig. 4 shows a preliminary result of interaction between the semantic (with randomly correlated patterns) and syntactic sub-networks. The detailed description of how we store the BLISS words and the grammar into the Potts network will appear elsewhere.

Table 1: Suggestion weights of some of the syntactic features, served as factors for the syntactic representation of words. Each element indicates the suggestion weight of a syntactic feature (labelled on the columns) for the generation of a word or words of a category (labelled on the rows).

	noun	verb	adj	conj	prep	pron	adv	sg	pl	neg	indf	def	propn
thatc	0	0	0	1	0	0	0	0	0	0	0	0	0
of	0	0	0	0	1	0	0	0	0	0	0	0	0
in	0	0	0	0	1	0	0.4	0	0	0	0	0	0
with	0	0	0	0	1	0	0	0.3	0	0	0	0	0
on	0	0	0.3	0	1	0	0.4	0	0	0	0	0	0
to	0	0	0	0	1	0	0.1	0	0	0	0	0	0
for	0	0	0	0.1	1	0	0	0	0	0	0	0	0
doesn't	0	1	0	0	0	0	0	1	0	1	0	0	0
don't	0	1	0	0	0	0	0	0.3	1	1	0	0	0
the	0	0	1	0	0	0	0	0.5	0.5	0	0	1	0
a	0	0	1	0	0	0	0	1	0	0	1	0	0
this	0	0	1	0	0	0.5	0	1	0	0	0	1	0
that	0	0	1	0	0	0.5	0	1	0	0	0	1	0
those	0	0	1	0	0	0.5	0	0	1	0	0	1	0
these	0	0	1	0	0	0.5	0	0	1	0	0	1	0
noun/sg	1	0	0	0	0	0	0	1	0	0	0	0	0
noun/pl	1	0	0	0	0	0	0	0	1	0	0	0	0
propn/sg	1	0	0	0	0	0	0	1	0	0	0	0	1
propn/pl	1	0	0	0	0	0	0	0	1	0	0	0	1
verb/sg	0	1	0	0	0	0	0	1	0	0	0	0	0
verb/pl	0	1	0	0	0	0	0	0	1	0	0	0	0
adjective	0	0	1	0	0	0	0.3	0	0	0	0	0	0



(a) Singular nouns

(b) Singular verbs

(c) Adjectives

Fig. 3: The average correlation $<Nas>$ of the syntactic representation of the words belonging to the same or different syntactic categories. (a) The correlation between the words that belong to singular nouns (Nsg) with themselves or with other word categories; likewise, for (b) singular verbs (Vsg), and (c) adjectives (Adj). The dashed lines indicate the expected correlations of randomly correlated patterns ($\simeq 3.2$, between content words).

church	trusts	holds	sweet	heavy	Zarathustra	holds	pearl	stands	heavy	Zarathustra,	pearl	stands	great
noun	verb		adj		[noun]	verb	noun	verb	adj	noun		verb	adj

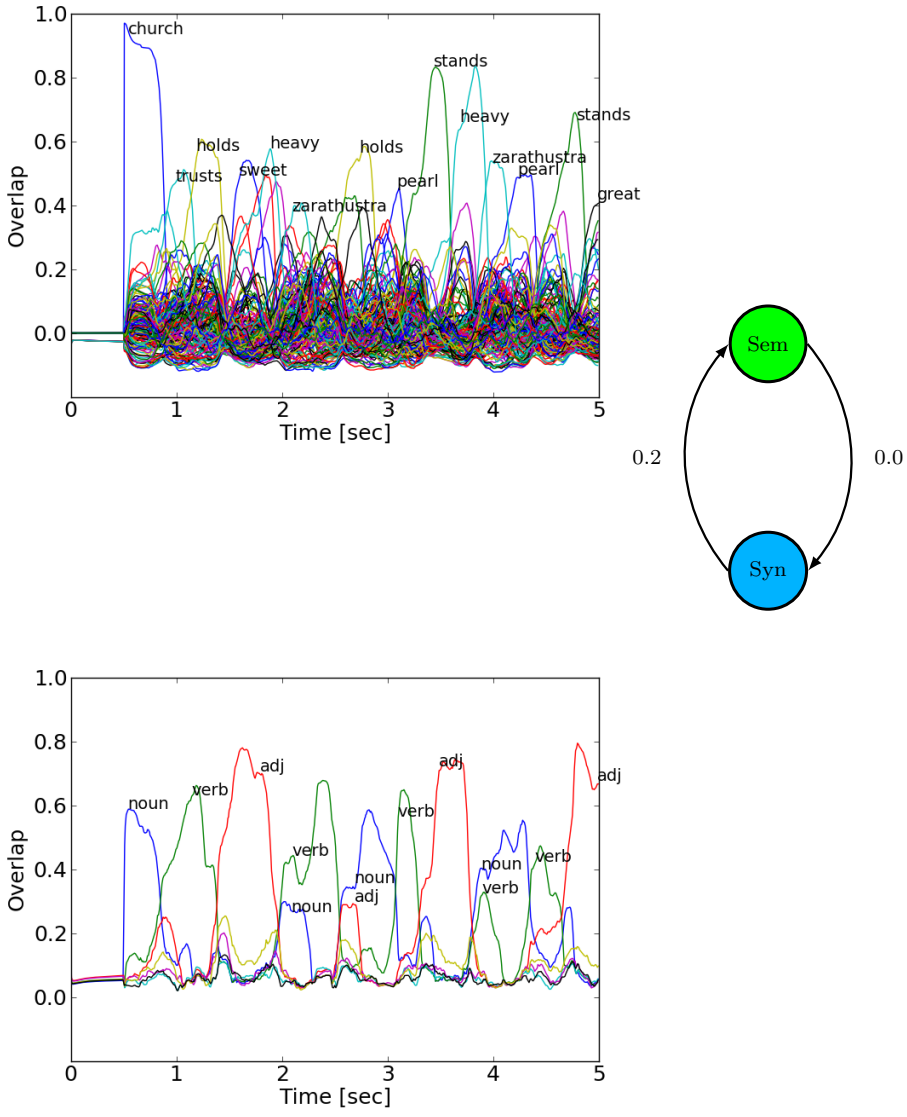


Fig. 4: The syntactic sub-network (bottom) influences the semantic sub-network (top) with the weight 0.2. The sentences produced by the interaction of these two sub-networks are written on the top. The parameters were set at $w = 1.6$, $\beta = 5$, $U = 0.0$; for the semantic sub-network, $C_{sem} = 100$, $\tau_{1\,sem} = 10$, $\tau_{2\,sem} = 200$, $\tau_{2\,sem} = 10000$; for the syntactic sub-network, $C_{syn} = 58$, $\tau_{1\,syn} = 5$, $\tau_{2\,syn} = 100$, $\tau_{3\,syn} = 5000$, $c_{auto} = 1.0$, and $c_{hetero} = 0.1$.

5. Discussion

We have encoded words of BLISS, our artificial language of intermediate complexity, into a Potts attractor neural network, a simplified model of the cortex with large storage capacity that includes two components, semantic and syntactic.

The distinction between semantic and syntactic representations of words is inspired by neuropyschological findings [28,29]. We have also made a distinction between the encoding of function words and content words, as suggested by several studies [30,31]. While we keep the overall activity for these two categories the same over the network, semantic units are less active for the function words than for the content words, while syntactic units are more active for the function words than for the content words.

By distributing a word on a network, we stayed away from extreme localized approaches in which the sentence constructs are represented on distinct set of units [1,3]; on the other hand, by having a sparse representation of the words, which are implemented as a set of features localized on Potts units, we did not follow extreme distributed approaches [5]. Further, by making a distinction between semantic and syntactic characteristics of a word, we embedded grammar knowledge in the Potts network, unlike the case of a simple recurrent neural network [10].

In spite of the considerations we gave for word representation, there remain limitations and future questions that need to be answered. A word, beside semantic and syntactic properties, is also associated to a sound structure, a property that needs to be considered in future representations of the words. The current implementation of BLISS, the training language of the network, does not contain pronouns, interrogative sentences, or embedding structure. To investigate the ability of the network to produce such sentences, one needs to first examine the length of dependences that the Potts network can handle. For randomly correlated patterns, the sequences stretch beyond a first-order Markov chain [35]; however, this measure needs to be investigated with sentences generated by the semantic and syntactic sub-networks, given that these sub-networks can be trained with different statistics of word transitions derived from BLISS corpora generated by the different semantics models.

References

1. J. Hummel and K. Holyoak, *Psychological Review* **104**, 427 (1997).
2. C. R. Huyck, *Cognitive Neurodynamics* **3**, 317(March 2009).

3. F. V. D. Velde and M. de Kamps, *Behavioral and Brain Sciences* **29**, 37 (2006).
4. T. Stewart and C. Eliasmith, Compositionality and biologically plausible models, in *The Oxford handbook of compositionality*, eds. M. Werning, W. Hinzen and M. Edouard (Oxford University Press, 2009)
5. R. W. Gayler, Vector symbolic architectures answer jackendoff's challenges for cognitive neuroscience, in *The Joint International Conference on Cognitive Science*, 2003.
6. T. Plate, Holographic reduced representations: Convolution algebra for compositional distributed representations, in *Proceedings of the 12th International Joint Conference on Artificial Intelligence*, 1991.
7. C. von der Malsburg, *Current opinion in neurobiology* **5**, 520(August 1995).
8. P. Smolensky, *Artificial Intelligence* **46**, 159(November 1990).
9. J. L. Elman, *Machine Learning* **7**, 195(September 1991).
10. J. L. Elman, *Cognition* **48**, 71(July 1993).
11. G. Borensztajn, The neural basis of structure in language: Bridging the gap between symbolic and connectionist models of language processing, ph.d. thesis, Institute for Logic, Language and Computation, University of Amsterdam2011.
12. S. Pirmoradian and A. Treves, *Cognitive Computation* **3**, 539 (2011).
13. G. Marcus, S. Vijayan, S. Bandi Rao and P. Vishton, *Science* **283**, 77 (1999).
14. R. Gomez, *Trends in Cognitive Sciences* **4**, 178 (2000).
15. J.-R. Hochmann, A. D. Endress and J. Mehler, *Cognition* **115**, 444 (2010).
16. M. Ullman, R. Pancheva, T. Love, E. Yee, D. Swinney and G. Hickok, *Brain Language* **93**, 185 (2005).
17. R. de Diego-Balaguer, L. Fuentemilla and A. Rodriguez-Fornells, *Journal of Cognitive Neuroscience* **23**, 3105 (2011).
18. J. Bahlmann, R. I. Schubotz and A. D. Friederici, *Neuroimage* **42**, 525 (2008).
19. J. Lany and J. Saffran, *Psychological Science* **21**, 284 (2010).
20. D. Amit, *Modeling Brain Function: The World of Attractor Neural Networks* (Cambridge University Press, 1992).
21. J. Hopfield, *Proceedings of the National Academy of Sciences* **79**, 2554 (1982).
22. I. Kanter, *Physical Review A* **37**, 2739(April 1988).
23. A. Treves, *Cognitive Neuropsychology* **22**, 276(January 2005).
24. C. Fulvi Mari and A. Treves, *Biosystems* **48**, 47(November 1998).
25. D. O'Kane and A. Treves, *Journal of Physics A: Mathematics and General* **25**, 5055 (1992).
26. V. Braitenberg and A. Schüz, *Anatomy of the Cortex: Statistics and Geometry.* (Springer-Verlag Publishing, 1991).
27. E. Kropff and A. Treves, *Natural Computing* **6**, 169(September 2006).
28. T. Shallice and R. Cooper, *The Organisation of Mind* (OUP Oxford, 2011).
29. K. Shapiro and A. Caramazza, The organization of lexical knowledge in the brain: The grammatical dimension, in *Cognitive neurosciences*, ed. M. Gazzaniga (Cambridge, MA: MIT Press, 2004) pp. 803–814, 3rd edn.
30. A. Friederici and P. Schoenle, *Neuropsychologia* **18**, 11 (1980).
31. a. D. Friederici, B. Opitz and D. Y. von Cramon, *Cerebral Cortex* **10**, 698(July

2000).
32. T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. a. Mason and M. A. Just, *Science* **320**, 1191(May 2008).
33. K. McRae, G. S. Cree, M. S. Seidenberg and C. Mcnorgan, *Behavior Research Methods* **37**, 547 (2005).
34. H. Sompolinsky and I. Kanter, *Physical Review Letters* **57**, 2861 (1986).
35. E. Russo, S. Pirmoradian and A. Treves, Associative latching dynamics vs. syntax, in *Advances in Cognitive Neurodynamics (II): Proceedings of the Second International Conference on Cognitive Neurodynamics*, 2011.