Scuola Internazionale Superiore di Studi Avanzati - Trieste

Towards Artificial Language Learning in a Potts Attractor Network

Candidate:

Sahar Pirmoradian

Advisor:

Alessandro Treves

Thesis submitted for the degree of Doctor of Philosophy in Neuroscience Area

Trieste, 2012

SISSA - Via Bonomea 265 - 34136 Trieste - ITALY

Contents

1 Introduction 1

2 BLISS: an Artificial Language for Learnability Studies 5

2.1 Basic Architecture of the Artificial Language . . . 7

2.1.1 Grammar . . . 8

2.1.2 Lexicon . . . 11

2.1.3 Semantics . . . 12

2.2 Tuning the Components to Produce Full Language Models . . . 12

2.2.1 Extraction of Statistics from Real Corpora . . . 12

2.2.2 Different Semantics Models . . . 13

2.3 Comparison among Model-generated Corpora . . . 19

2.3.1 Kullback-Leibler Divergence . . . 20

2.3.2 Markov Properties . . . 20

2.3.3 Adjusting Word Frequencies . . . 21

2.4 Results . . . 22

2.4.1 Distance Between Models: a Mild Effect of Semantics . . . 22

2.4.2 Mutual Information Between Words: a Mild Effect of Semantics . . . 25

2.4.3 Memory Characteristics . . . 27

2.5 Discussion . . . 29

2.6 Conclusion . . . 33

2.7 Appendix . . . 33

2.7.1 Full Grammar of BLISS . . . 33

2.7.2 Proof of Eq. 2.12 . . . 34

2.7.3 Length of BLISS Sentences . . . 34
2.7.4 Size of Model-Generated Corpora . . . 35

3 Encoding BLISS Words into a Potts Attractor Network 39
3.1 Evidence from the Brain for Word Representation . . . 39
3.2 Potts Attractor Network: a Simplified Model of the Cortex . . . 42
3.2.1 The Model . . . 44
3.3 Implementation of Word Representation in the Potts Network . . . 51
3.3.1 Generating Algorithm . . . 52
3.3.2 Semantic Representation . . . 54
3.3.3 Syntactic Representation . . . 56
3.3.4 The Potts Network stores the BLISS Words . . . 58
3.4 Results . . . 61
3.4.1 Word Correlations in the Semantic Representation . . . 61
3.4.2 Word Correlations in the Syntactic Representation . . . 66
3.5 Discussion . . . 69

4 The Potts Network Utters BLISS Words 73
4.1 Effect of the Network Parameters . . . 73
4.1.1 The Fixed Threshold: U . . . 74
4.1.2 The Self-reinforcement Term: w . . . 75
4.1.3 The Inverse Temperature: β . . . 76
4.1.4 The Time Constants: τ_1, τ_2, τ_3 . . . 77
4.2 The Potts Semantic Network . . . 79
4.3 The Potts Syntactic Network . . . 80
4.3.1 Latching with the Auto-associative Learning Rule Alone . . . 80
4.3.2 Turning on the Hetero-associative Learning Rule . . . 81
4.4 Interaction of Semantic and Syntactic sub-networks . . . 82
4.4.1 The Semantic sub-network with Correlated Patterns . . . 83
4.4.2 The Semantic sub-network Influences the Syntactic sub-network . . . 85
4.4.3 The Syntactic sub-network Influences the Semantic sub-network . . . 85

4.4.4 The Semantic sub-network and the Syntactic sub-network Co-operate 92
4.5 Discussion 95
4.5.1 Future Directions 96

5 Conclusion **99**

List of Figures

2-1 An example of the Exponential model 15

2-2 An example of the Subject-Verb model 17

2-3 An example of the Verb-Subject vs. the Subject-Verb model 18

2-4 An example of the No-Semantics model 19

2-5 KL-distances between word distributions 24

2-6 KL-divergence between the semantics models and the No-Semantics model 25

2-7 The mutual information words convey about other words in the same sentence 26

2-8 Markov properties of the Subject-Verb model 28

2-9 Memory effects in the Subject-Verb model 29

2-10 Probability distribution of sentence lengths 35

2-11 Average mutual information vs. number of sentences 36

2-12 Average triple mutual information vs. number of sentences 37

3-1 Conceptual derivation of the Potts network 44

3-2 An example of latching dynamics in the Potts network 45

3-3 The latching phase diagram 51

3-4 Order of word generation in the semantic sub-network 55

3-5 Order of word generation in the syntactic sub-network 57

3-6 Interaction between semantic and syntactic sub-networks 60

3-7 Frequency distributions of correlations in the semantic representation and randomly correlated patterns 62

3-8 The average correlation between the words within a semantic noun category 63

3-9 The correlation between words and their factors in their semantic representation vs. their joint probabilities in the BLISS corpus 64

3-10 The average correlation between a word category with other word categories in the semantic representation 65

3-11 The average correlation of the 7 lexical categories, the generating factors for the syntactic representation of words 67

3-12 The average correlation of the syntactic representation of the words belonging to the same or different syntactic categories 68

4-1 Effect of varying the fixed threshold, U, in the Potts network 75

4-2 Effect of varying the self-reinforcement term, w, in the Potts network . . . 76

4-3 Effect of varying the inverse temperature, β, in the Potts network 77

4-4 Effect of varying the time constants of the dynamic thresholds, τ_1, τ_2, and τ_3, in the Potts network 78

4-5 Limit cycle dynamics with correlated patterns 79

4-6 Dynamics of the syntactic sub-network with the slower and faster time constants 81

4-7 Hetero-associative learning introduces syntax 82

4-8 The syntactic sub-network influences the semantic sub-network with correlated patterns 84

4-9 The semantic sub-network influences the syntactic sub-network with a loose connection 87

4-10 The semantic sub-network influences the syntactic sub-network with a stronger connection 88

4-11 The syntactic sub-network influences the semantic sub-network with a loose connection 89

4-12 The syntactic sub-network influences the semantic sub-network with a stronger connection 90

4-13 Further improvement with slower semantics dynamics 91

4-14 A bidirectional interaction between the semantic and syntactic sub-networks with weak connections 93
4-15 A bidirectional interaction between the semantic and syntactic sub-networks with stronger connections 94

List of Tables

2.1 Samples of the BLISS PCFG and of its lexicon 10

2.2 Lexical categories, with examples 10

2.3 Samples of Sentences generated by the Subject-Verb model 23

2.4 The full BLISS PCFG 38

3.1 Suggestion weights of the syntactic features 72

Acknowledgements

I am deeply grateful to my advisor Alessandro Treves. This thesis would not have been possible without his encouragement, guidance, and support. I owe my sincere appreciation to him for helping me learn how to approach fundamental questions as a scientist and not as an engineer. I learned to playfully examine a phenomenon, excitedly observe and record its behaviour, and strictly set aside my prejudgement. I learned that the behaviour of the phenomenon must guide my intuition, and that I must not let my intuition guide the observation of the behaviour of that phenomenon. In some moments, I truly felt like a curious child, full of joy. I owe Ale for such valuable moments and lessons.

I would like to sincerely thank Sara Solla and Rens Bod for our fruitful discussions and for generously accepting to be on my defence committee. Sara and Rens have greatly helped me improve my thesis, as they read it with a great care and have provided me with thoughtful questions and helpful suggestions.

I should also thank Ritwik Kulkarni, who has been a cooperative colleague, a patient officemate, and an amazing friend for the past years. I will never forget the moments that we read papers together, discussed our scientific findings, and talked about many philosophical, social, and political issues. I also thank Giovanni Novembre, my other officemate, with whom I always shared my great awe for the beautiful view of the Adriatic sea we have had from our office in SISSA.

During my PhD I had a chance to meet many brilliant scientists who have greatly motivated me by their genius questions and scientific passion. My special thanks go to Giosue Baggio, Markus Muller, John Nicholls, Marina Nespor, Giuseppe Longobardi, John Hertz, Ulf Grenander, Matteo Marsili, Edmund Rolls, Ichiro Tsuda, Takashi Hashimoto, Francesco Battaglia, Jelle Zuidema, Cristian Micheletti, and Giuseppe Mussardo.

I am truly honored for having the emotional and intellectual support of my amazing friends throughout my PhD. My special thanks go to Daniela Saderi, Fahimeh Baftizadeh, Zhale Ghaemi, Ladan Amin, Shima Talehy, Shima Seyed-Allaei, Maryam Tavakoli, Silvia Benavides, Ana Laura, David Gomez, Montse Sole, Nahid Taherian, Hoda Alemi, Narges Javaheri, Katja Abramova, Gideon Borensztajn, Khadijeh Najafi, Francesca Pietracaprina, Fatemeh Rezvan, and Georgette Argiris. I also owe my gratitude to the SISSA staff, specially, Andrea Sciarrone, Alessio Isaja, Riccardo Iancer, and Federica Tuniz for their administrative and technical support.

I am sincerely grateful to my husband Alireza Alemi, whose emotional and intellectual support has been with me in every moment of the past four years. Alireza's insightful suggestions and our fruitful discussions were invaluable to me, especially during the last year when my supervisor was absent. No words can begin to express my heartfelt thanks to my parents Fatemeh Bujari and Sohrab Pirmoradian for their warm emotional support in my entire life. I sincerely thank them for their faith in me and allowing me to be as ambitious as I wanted. My brothers Soheil and Reza deserve my wholehearted thanks as well.

1

Introduction

"Church holds strong Zarathustra."

– Potts

It remains a mystery how children acquire natural languages; languages far beyond the few symbols that a young chimp struggles to learn, and with complex rules that incomparably surpass the repetitive structure of bird songs. How should one explain the emergence of such a capacity from the basic elements of the nervous system, namely neuronal networks?

To understand the brain mechanisms underlying the language phenomenon, specifically sentence construction, different approaches have been attempted to implement an artificial neural network that encodes words and constructs sentences (see e.g. (Hummel and Holyoak, 1997; Huyck, 2009; Velde and de Kamps, 2006; Stewart and Eliasmith, 2009)). These attempts differ on how the sentence constituents (parts) are represented—either individually and locally, or in a distributed fashion—and on how these constituents are bound together.

In *LISA* (Hummel and Holyoak, 1997), each sentence constituent (either a word, a phrase, or even a proposition) is represented individually by a unit—intended to be a population of neurons (Hummel and Holyoak, 2003)—and relevant constituents synchronously get activated in the construction of a sentence (or the inference of a proposition). Considering the productivity of the language—the ability of humans to create many possible sentences out of a limited vocabulary—this representation results in an exponential growth

in the number of units needed for structure representation.

In order to avoid this problem, *Neural Blackboard Architectures* (Velde and de Kamps, 2006) were proposed as systems endowed with dynamic bindings between assemblies of words, roles (e.g. *theme* or *agent*), and word categories (e.g. nouns or verbs). A neural blackboard architecture resembles a switchboard (a blackboard) that wires sentence constituents together via circuits, using highly complex and meticulously (unrealistic) organized connections.

As opposed to localized approaches, in a *Vector Symbolic Architecture* (Gayler, 2003; Plate, 1991), words are represented in a fully distributed fashion on a vector. The words are bound (and *merged*) together by algebraic operations—e.g. tensor products (Smolensky, 1990) or circular convolution (Plate, 1991)—in the vector space. In order to give a biological account, some steps have been attempted towards the neural implementation of such operations (Stewart and Eliasmith, 2009).

Another distributed approach was toward implementing a simple recurrent neural network that predicts the next word in a sentence (Elman, 1991). Apart from the limited language size that the network could deal with (Elman, 1993), this system lacked an explicit representation of syntactic constituents, thus resulting in a lack of grammatical knowledge in the network (Borensztajn, 2011; Velde and de Kamps, 2006).

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 first constructing an artificial language of intermediate complexity and then implementing 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 inspired manner on two components: one semantic and one syntactic.

As the training language of the network, we have constructed *BLISS* (Pirmoradian and Treves, 2011), a scaled-down synthetic language of intermediate complexity, with about 150 words, 40 production rules, and a definition of semantics that is reduced to statistical dependence between words. In Chapter 2, we will explain the details of the implementation

of BLISS.

As a sentence production model, we have implemented a Potts attractor neural network, whose units hypothetically represent patches of cortex. The choice of the Potts network, for sentence production, has been mainly motivated by the *latching* dynamics it exhibits (Kropff and Treves, 2006); that is, an ability to spontaneously hop, or latch, across memory patterns, which have been stored as dynamical attractors, thus producing a long or even infinite sequence of patterns, at least in some regimes (Russo and Treves, 2012). The goal is to train the Potts network with a corpus of sentences in BLISS. This involves setting first the structure of the network, then the generating algorithm for word representations, and finally the protocol to train the network with the specific transitions present in the BLISS corpus, using both auto- and hetero-associative learning rules. In Chapter 3, we will explain the details of the procedure we have adapted for word representation in the network.

The last step involves utilizing the spontaneous latching dynamics exhibited by the Potts network, the word representation we have developed, and crucially hetero-associative weights favouring specific transitions, to generate, with a suitable associative training procedure, sentences ”uttered” by the network. This last stage of spontaneous sentence production by the network has been explained in Chapter 4.

2

BLISS: an Artificial Language for Learnability Studies

"A noble sword fights for Ahuramazda."

– BLISS

In cognitive science, Artificial Language Learning (ALL) has become a prevailing paradigm for studying language acquisition (Saffran et al., 1996; Christiansen, 2000; Pena et al., 2002; Petersson et al., 2010). ALL involves training human adults (Friederici et al., 2002) and sometimes infants (Gomez, 2000) on artificial languages with particular structural constrains, and then testing their acquisition of structural information about the miniature languages they were briefly exposed to. The complexity of natural languages makes it difficult to isolate factors critical in language learning; consequently, artificial languages are used, instead, to control more precisely the environment of language learning and study selected aspects of natural languages.

ALL seems to be a valid tool for investigating natural language acquisition and processing. Several studies of event-related potential (Mueller et al., 2009) and neuroimaging (Petersson et al., 2010) have shown similar brain activation in both artificial language and natural language processing.

In addition, similar ALL paradigms have been administered to synthetic agents, such as artificial intelligence expert systems or connectionist networks (Kinder and Lotz, 2009),

which could not be possibly confronted with the daunting complexity of a natural language.

Artificial languages applied in both cognitive and computational studies, however, are usually designed on an *ad hoc* basis, and they typically include so few of the elements of a natural language, e.g. so few words and syntactic categories, as to make generalizations and the test of wide-ranging hypotheses entirely dependent on one's theoretical assumptions. In some studies, toy grammars are used to generate structured sequences without any reference to even the most basic structural patterns of natural languages (the controversial language *universals*) (Reber, 1967; Knowlton and Squire, 1996); while in others, which may include basic syntactic categories and phrase structure rules (Friederici et al., 2002; Opitz and Friederici, 2003), the rules and categories are so few as to raise questions as to how the properties investigated scale up in larger and more complex systems, which e.g. require the interaction of properties hitherto studied separately.

The aim of the present study is the construction of an artificial language, called *BLISS*, for Basic Language Incorporating Syntax and Semantics, which mimics natural languages by possessing 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 different language models. The term *language model* usually refers, in natural language processing contexts, to a model for predicting the next word given previous words (Manning and Schuetze, 1999). In this study, the number of degrees of freedom, in this sense, is effectively reduced by using the grammar, which groups words into lexical classes and thus restricts the domain of the next word.

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 (Marcus et al., 1999; Gomez, 2000), the special role of function words (Hochmann et al., 2010), the role of suffixes (Ullman et al., 2005), and especially hierarchical structures (de Diego-Balaguer et al., 2011; Bahlmann et al., 2008)

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 (Lany and Saffran, 2010), 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.

In contrast to small artificial languages that mostly focus on a single issue, BLISS provides us with several features of natural languages, including an interaction between syntax and semantics, function words, some nested structures, and at the same time BLISS can generate in no time corpora with representative sampling, avoiding at least in principle the problem of limited sampling of linguistic structures.

The structure of this chapter is as follows: the next section introduces the language components of BLISS: grammar, lexicon, and semantics. After that, we describe the way we fine tune the syntactic, lexical, and semantics components by deriving statistics from real corpora, thus defining different language models. Then, the methods used to compare the corpora generated by different language models are explained. Finally, we report the results obtained from the comparison between language models.

2.1 Basic Architecture of the Artificial Language

The architecture of BLISS is composed of a formal grammar, lexicon, and semantics. A formal grammar is a set of production rules for generating sentences from the vocabulary of the language. The lexicon of a language is its vocabulary, including its words and expressions. Though denoting a range of ideas, the term ”semantics”, in this study, focuses on statistical dependence between words, beyond that due to production rules. In this section,

we describe how these three components were designed in BLISS.

2.1.1 Grammar

The grammar assigned to BLISS is a Probabilistic Context-Free Grammar (PCFG). A PCFG is a 5-tuple G=(N, Σ, P, S, D). In this tuple, N is a set of nonterminal symbols that indicate phrasal categories (e.g. verb phrase) and lexical categories (e.g. verb). Σ is a set of terminal symbols which refer to the vocabulary (e.g. *sword*). P is a set of production rules, each of the form $A \rightarrow \beta$, where A is a nonterminal and β is a string of symbols from the infinite set of strings $(\Sigma \cup N)^*$, where $\cup$ denotes *set union* and $*$ is the *Kleene star*, used as an operator to concatenate zero or more strings from a given set (namely, $(\Sigma \cup N)$). S is a designated start symbol, and D is a function which assigns probabilities to each rule in P, expressing the probability that the given nonterminal A will be expanded to the sequence β; it is often denoted as $P(A \rightarrow \beta)$ or as $P(A \rightarrow \beta|\beta)$, where the sum of the probabilities of all possible expansions of a nonterminal is equal to 1 (Jurafsky and Martin, 2000).

The choice of the PCFG model has mainly been motivated by its simplicity and its context-independence assumption. PCFGs are the simplest model for tree structures—useful for capturing the complex recursive structure of natural languages—and they form the backbone of any more complex model (such as probabilistic tree-substitution grammar). The context-independence assumption is useful for our study as we have aimed to separately introduce the contextual information into BLISS for having a better control.

However, with the strong independence assumptions of the PCFG, two relatively separate sources of contextual information are ignored (Manning and Schuetze, 1999): lexical context and structural context. The structure probability of a sentence is influenced by the words in a sentence; for example, the chance of a verb phrase expanding as a verb followed by two noun phrases is much more likely with ditransitive verbs like *tell*, than with other verbs. To address this simple problem (among many other problems that the lack of the lexical information can raise), we incorporated some lexical information into the BLISS grammar by defining different verb types (e.g. *transitive*, *intransitive*, ...). PCFGs are also deficient on purely structural grounds, as they ignore the dependence on the relative

position in a parse tree; for instance, the probability of a noun phrase expanding in a certain way is independent of its position in the tree.This assumption is wrong, as for example in English, Pronouns and proper nouns appear much more commonly in subject position than in object position. These limitations have been addressed in other statistical models (e.g. *lexicalized PCFGs* (Charniak, 1995) or *left-corner parsers* (Rosenkrantz and Lewis, 1970)), that we do not use in the current implementation of the BLISS for simplicity.

In the Chomsky hierarchy (Chomsky, 1959), a context-free grammar, as opposed to a regular grammar, can have center-embedded recursion (of the form $\mathtt{P}(A \rightarrow \alpha A \beta)$), such as relative clauses within a sentence. Context-free languages with center-embedded recursion cannot be generated by a finite state automaton (Chomsky, 1959). In the BLISS grammar, for simplicity, the only recursion structure is *that-complement-clause* preceded by embedding verbs such as *think, believe, know,* etc. Obviously the grammar could easily be extended to have more complex syntactic structure.

The BLISS PCFG in its current implementation (v1.1) consists of 39 nonterminal symbols, 150 terminal symbols, and 40 production rules with their assigned probabilities (See Table 2.1 for samples and Table 2.4 in Appendix for the full grammar). Nonterminal symbols, represented by <.>, include the start symbol (S), singular and plural noun phrases (NP1/NP+), verb phrases (VP1/VP+), determiner phrases (DP1/DP+), prepositional phrases (PP), nouns (Noun1/Noun+), verbs (Verb1/Verb+), adjectives (Adj), prepositions (Prep), and a complementizer (ThatClz); each of which was subcategorized as listed in Table 2.2. The procedure of selecting terminal symbols is discussed in section Lexicon. The way we enumerated the production rules is indicated in Table 2.1.

Further, inflectional suffixes used in the grammar are: *-s* plural (e.g. *swords*), *-s* third person singular present (e.g. *kills*), and *-n't* negative (e.g. *don't*).

Using the context freeness assumption and the chain rule, it follows that the probability of a a particular derivation sequence der derived by a PCFG is defined as the product of the probabilities of all the rules r used to expand each left-hand-side nonterminal in the derivation sequence (Jurafsky and Martin, 2000):

Table 2.1: Samples of the BLISS PCFG and of its lexicon.

No.		**Rule**	**Probability**
1	<S>	→ <DP1><VP1>	0.50
2		→ <DP+><VP+>	0.50
3	<VP1>	→ <Verb1>	0.85
4		→ <Neg1><Verb1>	0.15
5	<Verb1>	→ <TransVerb><DP>	0.37
6		→ <IntransVerb1>	0.41
7		→ <EmbedVerb1><S>	0.07
8		→ ...	...
9	<Noun1>	→ *sword*	0.13
		→ *dog*	0.05
		→ ...	...
10	<TransVerb1>	→ *kills*	0.03
		→ *guards*	0.02
		→ ...	...

Table 2.2: Lexical categories, with examples. The last column indicates the number of words of the corresponding subcategory which are used in the BLISS grammar.

Lexical Category	**Subcategories**	**Nonterminal Symbols**	**Words**	**No.**
Verbs	intransitive	<IntransVerb1>	*comes, dies, ...*	8
	monotransitive	<TransVerb1>	*loves, guards, ...*	20
	ditransitive	<DitransVerb1>	*gives, brings*	2
	embedding	<EmbedVerb1>	*wishes, believes, ...*	7
Nouns	common	<Noun1>	*sword, dog, ...*	18
	proper	<PropNoun1>	*Zarathustra, Ahuramazda, ...*	4
Determiners	article	<Det1>	*the, a*	2
	demonstrative	<Dems1>	*this, that*	2
Prepositions	following nouns	<PP-n>	*of, in, with, on*	4
	following ditransitive verbs	<PP-vb >	*to*	1
	following verbs	<PP-to>	*for, in, with*	3
Complementizers		<ThatClz>	*that*	1
Adjectives		<Adj>	*great, noble, ...*	18

$$P(der) = \prod_i P(r_i). \tag{2.1}$$

The probability of a sentence produced by a PCFG is the sum of the probabilities of all the possible derivations that generate that sentence:

$$P(S) = \sum_{der} P(der, S) = \sum_{der} P(der)P(S|der) = \sum_{der} P(der); \tag{2.2}$$

in the BLISS grammar, there is only one possible derivation for each BLISS sentence.

In a derivation sequence producing a sentence, there are some production rules which contain only nonterminals on their right hand side (excluding the nonterminals that expand to the terminals or words). We define the collection of these production rules as the underlying structure that generate the corresponding sentence, and we denote it as str, and the probability of this sequence derived by grammar G is denoted as $P(str)$ or more precisely as $P(str|G)$.

2.1.2 Lexicon

The BLISS lexicon contains about 150 lexical words, including singular and plural forms. The words were extracted from the Shakespeare corpus, which includes about 1,000,000 word *tokens*, as multiple occurrence of about 27000 word *types* (vocabulary words). The BLISS lexicon contains different category of words such as nouns, verbs, adjectives (content words), and prepositions, articles, demonstratives, complementizer (function words), as listed in Table 2.2. The content words (except for *proper* nouns) were selected from a set of high frequency words in the Shakespeare corpus. The selection was unique, regardless of either singular or plural form. For *common* nouns, after extracting frequent nouns from Shakespeare, we chose among them some which were shared with a database (McRae et al., 2005), in which a set of feature norms were collected for 541 living and non-living basic-level concepts. Knowing the feature norms of BLISS nouns enables us to derive distributional statistics, when desirable, from these norms, such as their pairwise correlation. Further, BLISS common nouns were chosen so as to be categorized into animates (e.g. dog, horse, . . .), buildings (e.g. church, house, . . .), and objects (e.g. sword, crown, . . .).

As proper nouns, we selected four singular proper nouns (Zarathustra, Ahuramazda, Ahriman, Yasmin) and two plural proper nouns (Magi, Greeks), inspired by Friedrich Nietzsche's *Thus Spake Zarathustra*.

2.1.3 Semantics

In BLISS, semantics is defined as the statistical dependence of each word on other words in the same sentence, purely determined by imposing constraints on word choice during sentence generation. After the grammar chooses a legal lexical category by production rules, the words of the chosen category compete for the appearance in a sentence on the basis of their relative frequency and joint probabilities, which were extracted from the Shakespeare corpus, following a procedure to be explained later.

2.2 Tuning the Components to Produce Full Language Models

After designing a general BLISS grammar and vocabulary of intermediate complexity, we aim to make it as natural as possible. Therefore, we adjust the transition probabilities of grammar rules, from nonterminals to either nonterminal or terminal symbols, as well as the joint frequency between pairs of words, to real corpora. We choose two corpora, the WSJ and Shakespeare. In this section the procedure followed for this adjustment is explained. Additionally, we elaborate on our definition of semantics and the difference between the four language models we introduce.

2.2.1 Extraction of Statistics from Real Corpora

In the BLISS grammar, the probabilities of transitions from nonterminal to nonterminal symbols, i.e. between phrasal and lexical categories, were adjusted to the statistics derived from the WSJ corpus[1], a syntactically annotated corpus, with about 1,000,000 word tokens and 38000 word types. Using tgrep2 software, we derived the frequency of all lexical categories which were used in the BLISS grammar and a probability was assigned to each category according to its derived frequency. The transition probability of nonterminal symbols to function words, i.e. prepositions, determiners, and auxiliary verbs, were adjusted to WSJ as well.

[1] Treebank-3 release of the Penn Treebank project

On the other hand, the transition probabilities from lexical categories to content words were extracted from the Shakespeare corpus, to give it a less prosaic feel. These probabilities of words are called *prior probabilities* in this paper, i.e. prior to the choice of other words in the same sentence. The same nouns and verbs as well as their probabilities were used for singular and plural categories (for example, see probabilities in the rules starting with <N1> and <N+> in Table 2.4). Proper nouns were set as equiprobable in BLISS (see <PropN1> and <PropN+> in Table 2.4).

Beside the frequency of single words (1st order statistics), the joint frequency of word-pairs — 2nd order statistics — was also extracted from the Shakespeare corpus. The joint frequency $f(w_i, w_j)$ was calculated for each pair of content words (nouns, verbs, adjectives); not for function words and not for Proper Nouns. $f(w_i, w_j)$ is the number of Shakespeare sentences in which the content words (w_i, w_j) orderly appear together consecutively or non-consecutively. Because of the poetic style in Shakespeare, having very long sentences with many short phrases, we counted the occurrence of a word-pair within a window of 5 words in a sentence. The probability of a word-pair $P(w_i, w_j)$ is calculated by dividing its frequency, which is extracted from the Shakespeare corpus, by the number of all the possible word-pairs (all pairs of the BLISS content words, excluding the proper nouns).

2.2.2 Different Semantics Models

After the extraction of the probability of grammar rules, of words, and of the joint frequencies from real corpora, we construct alternative language models based on the different selection algorithms applied for choosing content words. After a legal lexical category is selected by the grammar, the words of the chosen category compete for the appearance in a sentence based on their *selection probability*. The *selection probability* of content word w_n at position n is calculated as a weighted sum of the prior probability of the word (the probability encoded in the grammar $\mathtt{G}$) and its *semantics probability*, which is the conditional probability of the word given the preceding words $w_1..w_i..w_{n-1}$ in the same sentence (denoted as the history h in Eq.2.3). The probability function P which is applied as the se-

lection probability of the next word in a sentence is

$$P(w_n|h,G) = (1-g) * P_{prior}(w_n|G) + g * P_s(w_n|h), \tag{2.3}$$

where

$$P_s(w_n|h) = \frac{1}{C}\left(c_1 P(w_n|w_1) + ... + c_{n-1} P(w_n|w_{n-1})\right), \tag{2.4}$$

and g is a *semantics coefficient*, which is set to zero when the selection of a word only depends on the grammar, with no dependence on other words in the sentence. Semantics can then be switched on and off by changing the parameter g.

In Eq. 2.3, P_{prior} and P_s denote prior probability and semantics probability, respectively. The c_i's are *dependence coefficients*, and $C = \sum_{i=1}^{n-1} c_i$; if $c_i \neq 0$, the word w_i is called a *semantically effective word*. The conditional probabilities $P(w_n|w_i)$ between content words were adjusted to the Shakespeare corpus in the manner explained in the previous section (and they are zero when either or both are function words). For example, if the word w_n belongs to the singular noun category (*nsg*), $\sum_{w_k \in nsg} P(w_k|w_i) = 1$, where the sum is over all competing words (including the word w_n) belonging to the singular noun category.

The probability of sentence $S = w_1...w_i...w_N$ whose word selections are influenced by both the grammar and history words reads:

$$P(S|h,G) = P(str|G) * \prod_{i=1}^{N} P(w_i|h,G), \tag{2.5}$$

where $P(str|G)$ is the probability of the underlying structure of the sentence (see section 2.1.1), and $P(w_i|h,G)$ is the selection probability of each word (Eq. 2.3).

Based on different choices of parameters in the semantics probability function (Eq. 2.4), 4 different language models, 3 with semantics and one without semantics, were defined: the *Exponential*, *Subject-Verb*, *Verb-Subject*, and *No-Semantics* models.

Exponential Language Model

In the Exponential model, all preceding content words in a sentence affect the next word; dependence coefficients in the semantics probability function (Eq. 2.4) are exponential

functions of the distance between preceding words and the current word, $c_i = e^{\lambda(i-n)}$; that is, words at the beginning of a sentence are less effective than the words which are close to the word currently being selected. Thus, the semantics probability function looks like

$$P_s(w_n|h) = \frac{1}{C}\left(e^{\lambda(1-n)}P(w_n|w_1) + ... + e^{\lambda(-1)}P(w_n|w_{n-1})\right) , \tag{2.6}$$

where $\lambda > 0$ is a *temporal decay coefficient*. An illustration is given in Figure 2-1a for the sentence *The bloody sword kills*, which according to the grammar is to be followed by a plural noun (Figure 2-1b); all words in this lexical category compete to be chosen. The *selection probability* of each word, e.g. *houses*, *swords*, *calves*, ..., is measured as the weighted sum of their prior probability (Figure 2-1b), and their semantics probability (Eq. 2.6). In Figure 2-1a, as $\lambda = 1/3$, since the distance between *kills* and the new word is 1, the dependence coefficient of the effective word *kill* is $e^{-(1/3)}$.

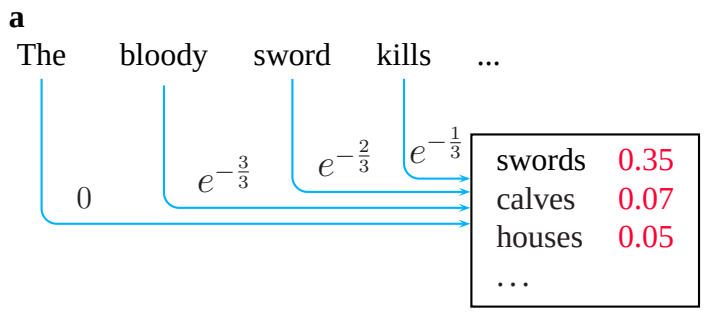


Figure 2-1: An example of the Exponential model. (a) The words of plural nouns compete for the appearance in the currently generated sentence *The bloody sword kills*. Their prior probabilities represented in (b) are modified by the effect of the preceding content words, with an exponential factor, and turned into the probabilities written in red in the box; the words that are highly affected by turning on the semantics ($g = 1.0$) are written. The conditional probabilities of words extracted from the Shakespeare corpus are $P(\text{swords}|\text{bloody}) = 0.38$, $P(\text{houses}|\text{bloody}) = 0.15$, $P(\text{calves}|\text{bloody}) = 0.0$; $P(\text{swords}|\text{sword}) = 0.28$, $P(\text{houses}|\text{sword}) = 0.05$, $P(\text{calves}|\text{sword}) = 0.0$; $P(\text{swords}|\text{kills}) = 0.38$, $P(\text{houses}|\text{kills}) = 0.0$, $P(\text{calves}|\text{kills}) = 0.15$.

Subject-Verb Language Model

In the Subject-Verb model, we presume the subject and verb of a sentence to quite strongly influence which word will come next. Therefore, the semantics probability is only affected by the subject and verb of a clause. The dependence coefficients of other words, other than the subject and verb of the clause, are zero:

$$\begin{aligned} P_s(w_n|h) \quad &= \\ & c_{subject}P(w_n|w_{subject}) + c_{verb}P(w_n|w_{verb}) \ , \end{aligned} \tag{2.7}$$

where $c_{subject} + c_{verb} = 1$ (the normalization factor, C, is omitted as $C = 1$). If there is still no verb in a sentence, $c_{subject} = 1$. Moreover, in case of a *that-complement-clause*, the subject and verb of the specific clause (either the main clause or the complement clause) to which the currently selected word belongs are those considered as effective words. In Figure 2-2, only the subject and verb of the sentence *The bloody sword kills*, i.e. *sword* and *kills* respectively, are effective, and in this example $c_{subject} = 0.2$, and $c_{verb} = 0.8$. As illustrated, the selection probabilities of the competing words (Figure 2-2a) are different from the prior probabilities of these words (Figure 2-2b).

When the subject and verb of a sentence are chosen, they influence not only the choice of successive words, but also the words previously produced in the same sentence. In the example, the word *bloody* might be another adjective, which replaces it as another word more likely to appear with the subject *sword* and the verb *kill*.

Verb-Subject Language Model

In the Verb-Subject model, only the subject and the verb affect the semantics probability of other words in the sentence or clause, same as in the Subject-Verb model, and with the same probability. However, in the Verb-Subject model, first a verb is chosen, and then a noun that is likely to appear with the chosen verb is selected as the subject. Figure 2-3 illustrates the difference between these two models. As shown in Figure 2-3a for the Verb-Subject model, first the verb *kills* was chosen, then a subject among the possible nouns, e.g. *house*, *sword*, *calf*, ...; whereas as shown in Figure 2-3b for the Subject-Verb model, first

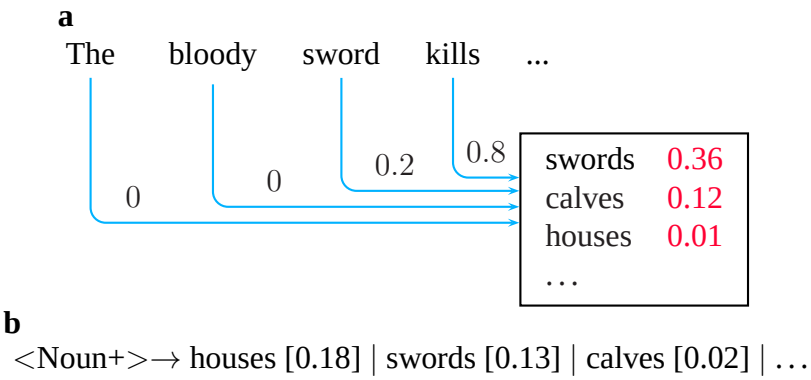


Figure 2-2: An example of the Subject-Verb model. (a) The words of plural nouns compete for the appearance in the currently generated sentence *The bloody sword kills*. Their prior probabilities represented in (b) are influenced only by the subject (*sword*) and the verb (*kills*) of the sentence and are changed to the probabilities written in red in the box. The subject and verb of a sentence constraint the choice of not only the successive words but also those already generated; that is, the word *bloody* might replace another adjective less probable to appear with *sword* and *kills*. The words that are highly affected by turning on the semantics ($g = 1.0$) are written in the box. The conditional probabilities of words extracted from the Shakespeare corpus are $P(\text{swords}|\text{sword}) = 0.28$, $P(\text{houses}|\text{sword}) = 0.05$, $P(\text{calves}|\text{sword}) = 0.0$; $P(\text{swords}|\text{kills}) = 0.38$, $P(\text{houses}|\text{kills}) = 0.0$, $P(\text{calves}|\text{kills}) = 0.15$.

the subject was chosen, i.e. *sword*, next the verb. As depicted in this figure, the selection probability of either noun (a) or verb (b) is different from their prior probability in (c) and (d), respectively.

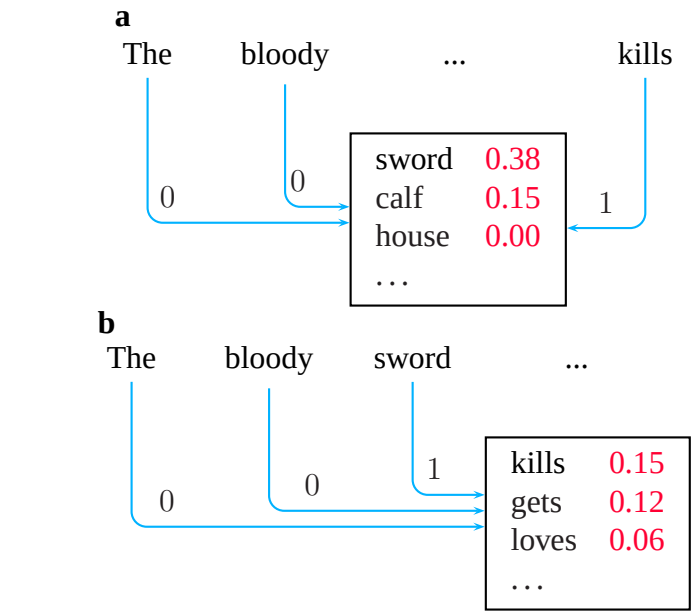
No-Semantics Language Model

In the No-Semantics model, the semantics is switched off, i.e. $g = 0$ in Eq. 2.3. As Figure 2-4a shows, preceding words do not have any effect on the selection probability of currently selected words, which is the same as the prior probability of these words (Figure 2-4b).

Generating Corpora without Grammar

As a control to measure the effect of grammar in BLISS, we produced corpora without following a grammar, although one of the semantics models explained earlier was adopted in one variant.

Unigram: Assuming entire independence between words of a sentence, we used the so-called Unigram language model, in which the selection probability of a word is its *relative*



c

<Noun+>→ house [0.18] | sword [0.13] | calf [0.02] | ...

d

<TransVerb1>→ loves [0.23] | kills [0.04] | gets [0.04] | ...

Figure 2-3: An example of the Verb-Subject vs. the Subject-Verb model. (a) In the Verb-Subject model, first the verb, i.e. *kills*, is chosen by its prior probability, *0.03* in (d). Then, the words of singular nouns compete with probabilities that are influenced by the verb, thus their prior probabilities (c) will change to the probabilities written in red in the box. (b) In the Subject-Verb model, first the subject, i.e. *sword*, is chosen by its prior probability, *0.13* in (c). Then the words of singular transitive verbs compete with probabilities that are influenced by the subject, so their prior probabilities (d) will change to the probabilities written in red in the box. In both of the models, the subject and verb influence the choice of not only the following words but also the preceding words, e.g. *bloody*. The words that are highly affected by turning on the semantics ($g = 1.0$) are written in each box. The conditional probabilities of words relevant for the Verb-Subject model are $P(\text{sword}|\text{kills}) = 0.38$, $P(\text{calf}|\text{kills}) = 0.15$, $P(\text{house}|\text{kills}) = 0.00$. The conditional probabilities of words relevant for the Subject-Verb model are $P(\text{kills}|\text{sword}) = 0.15$, $P(\text{gets}|\text{sword}) = 0.12$, $P(\text{loves}|\text{sword}) = 0.06$.

frequency in a corpus generated by the No-Semantics model; the relative frequency of word w_n is the number of times the word appears in the corpus, $C(w_n)$, divided by the total number of word tokens N in the entire corpus:

$$P(w_n|h, G) = P(w_n) = \frac{C(w_n)}{N}; \tag{2.8}$$

both the preceding words, h, and the grammar, G, are disregarded in the selection of the

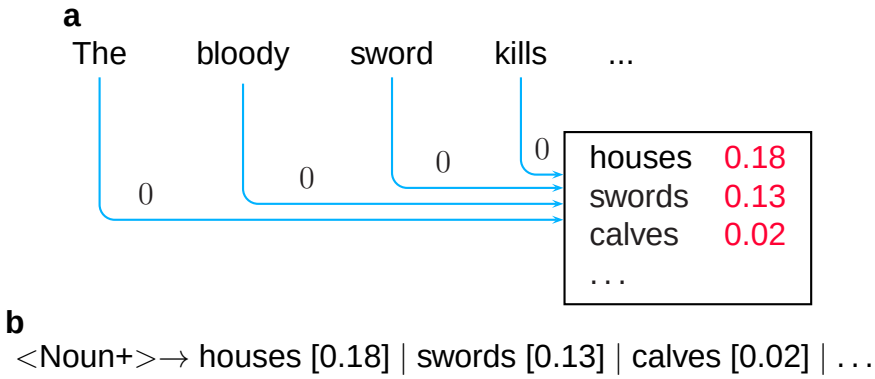


Figure 2-4: An example of the No-Semantics model. (a) The words of plural nouns compete by their prior probabilities in (b), not influenced by the preceding words.

word.

Exp-NoSyntax: To measure merely the effect of semantics, in the absence of grammar, we defined an Exp-NoSyntax model in which the semantics effect (in Eq. 2.3) is $g = 1$ and the dependence coefficient is defined as in the Exponential model (Eq. 2.6):

$$\begin{aligned} & P(w_n|h, G) \\ & \quad = P_s(w_n|h) \\ & \quad = \tfrac{1}{C}\left(e^{\lambda(1-n)}P(w_n|w_1) + ... + e^{\lambda(-1)}P(w_n|w_{n-1})\right) . \end{aligned} \tag{2.9}$$

A Further Control Model

Equiprobable: As a control for the No-Semantics model, we also introduced the Equiprobable model, which is the same as the No-Semantics one, except for the prior probability of words. Words which belong to the same lexical category are equiprobable, not following the probabilities derived from Shakespeare. For instance, the probability of plural demonstratives (<Dem+> in the last row of Table 2.4) changes from 0.60 and 0.40, for *these* and *those*, respectively, to 0.50 and 0.50.

2.3 Comparison among Model-generated Corpora

To compare corpora generated by the different semantics models, we used methods of Information Theory, with the necessary controls to have reliable measures. The effect of

introducing semantics into BLISS was measured as the distance between distributions of word-pairs in corpora generated by the different language models.

Before comparing word-pair distributions, one should eliminate the effect of word frequency differences, so as to measure purely semantics, i.e. word-dependence effects. That is, only corpora which have the same or very close word frequencies can be compared to assess semantics effects, which then result in differences in pair distributions. In this section, first we detail the methods used for measuring semantics effects. Further, we describe how the word frequencies of the semantics models were adjusted for comparison to the No-Semantics model.

2.3.1 Kullback-Leibler Divergence

The Kullback-Leibler (KL) divergence (or distance) is a non-symmetric measure of the difference between two distributions (Cover and Thomas, 1991); it measures how many extra bits are required to code a random variable having a true probability distribution P using a code based on an assumed and not correct probability distribution Q:

$$D_{KL}(P||Q) = \sum_{w} P(w) \log_2 \frac{P(w)}{Q(w)}. \tag{2.10}$$

This measure is applied here to compare the differences between word distributions, as well as word-pair distributions, of corpora generated by the different language models. Although non-symmetric, i.e. $D_{KL}(P||Q) \neq D_{KL}(Q||P)$, the asymmetry is small for the distributions considered here, almost always in the third decimal place. In the Result section, we show the average of the two KL-distances, rounded to second decimal precision.

2.3.2 Markov Properties

A stochastic process has the Markov property if the present state predicts future states independent of how the current state is obtained; that is, the process is memory-less. Considering BLISS as a stochastic process, we aim to see if it has the Markov property; if, as expected, it does not, we examine how much it deviates from a Markov process, that is,

from a (first order) Markov chain.

To quantify how close BLISS is to a first order Markov chain, we measure the three-way mutual information (MacKay, 2003), defined as:

$$I(X_n; X_{n-2}, X_{n-1}) = I(X_n; X_{n-1}) + I(X_n; X_{n-2}|X_{n-1}), \tag{2.11}$$

where X_{n-2}, X_{n-1}, and X_n are the values of a random variable in three consecutive states, and $I(X_n; X_{n-2}, X_{n-1})$ is the amount of information the values of this random variable in the first and second states, $n-2$ and $n-1$, convey about the variable in the third state, n.

A random variable has a first order Markov property if its distribution only depends on the distribution of the previous state, i.e., if in Eq. 2.11, $I(X_n; X_{n-2}|X_{n-1}) = 0$. We measure this quantity in the corpora generated by our models, where X_i is the random variable for position i in the sentence and takes as value the words of the vocabulary. From now on, we use $I(i; i-1)$ instead of $I(X_i; X_{i-1})$, for simplicity, when referring to a mutual information value relative to a specific position i, and $I(n; n-1)$ when referring to its average across positions.

2.3.3 Adjusting Word Frequencies

As explained at the beginning of this section, we need to eliminate the effect of first-order statistics, i.e. word frequencies by producing corpora with close word frequencies; that is, corpora must be generated with pre-assigned posterior word frequencies. This goal can be achieved by appropriately adjusting the *prior* word frequencies used by each language model.

We analysed the relation between the prior probability of a word w_j and its posterior probability after it is observed having a specific role r (i.e. subject, verb, ...) in sentences generated by the BLISS language models (See Eq. 2.13 in Appendix for details):

$$P_{prior}(w_j) = \frac{P_{post}(w_j|r) - g\sum_{h_e} P(h_e)P_s(w_j|h_e)}{\sum_{h_{ie}} P(h_{ie}) + (1-g)\sum_{h_e} P(h_e)}, \tag{2.12}$$

where all possible configurations of preceding words (history h) are considered as either *effective* history (h_e) or *ineffective* history (h_{ie}). A history is called effective if $P_s(w_j|h) \neq 0$, and ineffective if $P_s(w_j|h) = 0$. Knowing $\sum_{h_{ie}} P(h_{ie}) + \sum_{h_e} P(h_e) = 1$, we see in this equation that if $g = 0$, then $P_{prior}(w_j) = P_{post}(w_j)$, for the No-Semantics model, where the role (r) of the word does not influence its probability $P(w_j|r) = P(w_j)$.

Using Eq. 2.12, prior probabilities of content words in the semantics models (the Subject-Verb, Verb-Subject, and Exponential) were calculated to yield the same posterior probabilities as those in the No-Semantics model. In this fashion, we generated corpora with very close word frequencies by the different language models. The distances between word distributions were measured using KL-divergences (see Results).

2.4 Results

Applying the methods described above, we have constructed three language models with semantics, as well as one language model with syntax but without semantics (No-Semantics) and one with semantics but without syntax (Exp-NoSyntax), to serve as controls, together with two straightforward controls, one with neither syntax nor semantics (Unigram) and one with syntax but no individual word frequency tuning (Equiprobable). Table 2.3 shows some examples of sentences in the Subject-Verb model.

To quantitatively assess the dependencies between words introduced by the various language models, we have measured KL-distances between the word-pair distributions of these models; the results are discussed in the first part of this section. In the second part, we focus on individual sentences to measure the amount of information conveyed by its constituent words. Finally, we investigate the memory characteristics of BLISS.

The length of the sentences and the size of the corpora needed for the measures used in this section is discussed in the Appendix.

2.4.1 Distance Between Models: a Mild Effect of Semantics

In order to compare word-pair distributions among models, first we need to make sure that individual word frequencies are almost equal in all the main models (No-Semantics,

Table 2.3: Samples of Sentences generated by the Subject-Verb model

The church stands.
The precious crown guards the royal sword.
Crowns give Zarathustra to the sword.
The sweet horse knows that a church believes that Ahriman dies.
A noble sword fights for Ahuramazda.
Zarathustra loves a holy church.
Rocks of bloody doors cut Yasmin.
That noble dagger keeps the house.
Horses don't go.
Foul gates enter.
The dogs don't fight.
A gracious dog doesn't come with the sheep.
A house of Zarathustra thinks that Ahriman keeps the doors.
A royal house sits.
Gates with a strong sword praise the horse.
The swords prove that Magi hold the wall.
The sword doesn't die.
A gate thinks that Ahriman guards a house in the door.
Strong rocks don't sit.
The dog doesn't wish that the gates in the holy house keep dogs.

Subject-Verb, Verb-Subject, and Exponential). This is because comparing second-order statistics (word-pair distribution) is meaningless unless the first-order statistics (individual word frequencies) are equal.

In Figure 2-5a, the KL-distances between individual word frequencies are shown, after adjusting the word frequencies for the semantics models to those in the No-Semantics one. The vertices of the pyramids indicate the language models. The number in (and, approximately, the size of) the circles on the edge between each two vertices indicates the KL-distance between the word distributions of corpora generated by the corresponding models. As shown, the distances between the No-Semantics and the three semantics models are zero, to 2nd decimal precision. As a control, the No-Semantics model is at a considerable distance from the Equiprobable model, and, likewise, the Exponential model is at a large distance from the equivalent exponential model without syntax (Exp-Nosyntax). The Exp-Nosyntax model is at about the same distance, in fact, from the other two semantics models (not shown).

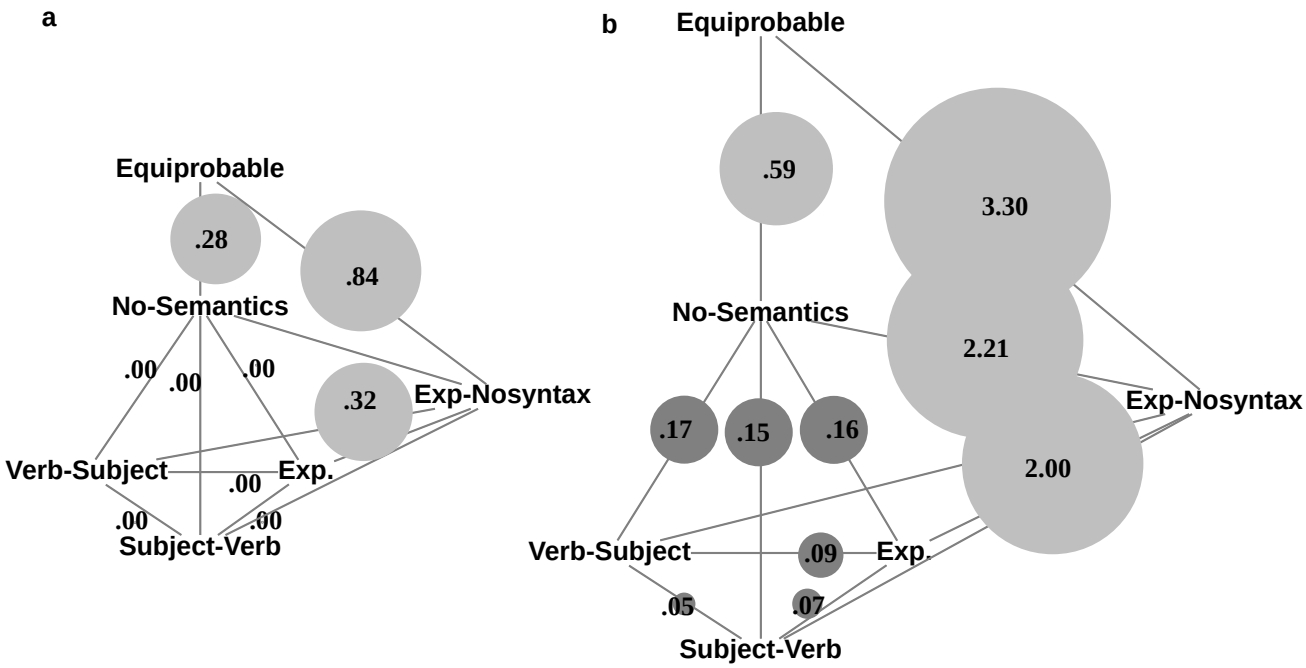


Figure 2-5: KL-distances between individual word distributions (a) and word-pair distributions (b). The vertices of these pyramids indicate the distributions of the No-Semantics, Subject-Verb, Verb-Subject and Exponential main language models, as well as the Equiprobable and Exp-Nosyntax control models. The number within and the size of the circles on the edge between each two vertices indicate the KL-distance between the distributions of the corpora generated by the corresponding models. The semantics coefficient was set at $g = 0.9$.

Given the same word frequencies across the main language models, we expect to see different word dependencies between the models, especially between the No-Semantics model and the semantics ones due to the semantics effect. We use the KL-divergence between word-pair distributions of every two models to quantify the difference between the word dependences of these two models: the greater the KL-divergence between the two word-pair distributions, the larger the difference in word dependencies between the two models.

In Figure 2-5b, adopting the same pyramidal graphic scheme as in Figure 2-5a, the KL-distances are shown between word-pair distributions among the same models. As illustrated, the No-Semantics model is at a considerable distance from each of the semantics models, larger than the distance between each pair of these semantics models. This result quantifies semantics effects in these models. Note that the effects of syntax on word-pair

distributions are much larger, as shown by the distances to the control models without syntax (the Exp-Nosyntax). The distance between the Equiprobable and the No-Semantics model demonstrates the necessity for adjusting word frequencies across models. These are in fact the same model (without a semantics effect) except for their word frequencies. Hence, the distance between pairwise statistics of these two models (Figure 2-5b) merely reflects the (large) effect of unequal word frequencies; such effect clearly needs to be removed to correctly assess differences arising from unequal word dependences.

Figure 2-6 shows that the distance between the semantics models and the No-Semantics one gets monotonically larger as we increase the parameter g which controls semantics effects (Eq. 2.4). Theoretically g could vary between 0 (when the semantics is switched off) and 1 (when the selection probability of a word is only influenced by the preceding history, not by its prior probability); however, for $g > 0.9$ it became impossible to generate corpora with the same posterior probability for single words, as per Eq. 2.12.

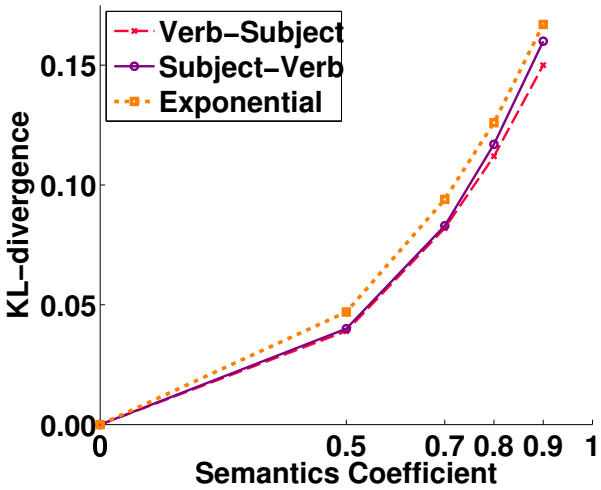


Figure 2-6: KL-divergence between the semantics models and the No-Semantics model, when the semantics parameter g changes. Note that g can only reach up to $g = 0.9$, before Eq. 2.12 becomes inapplicable for adjusting word frequency.

2.4.2 Mutual Information Between Words: a Mild Effect of Semantics

To find out how much information the words in a sentence convey about the other words in the same sentence, that is, to what extent successive word choices are mutually constrained, we looked at different measures of mutual information (see Figure 2-7).

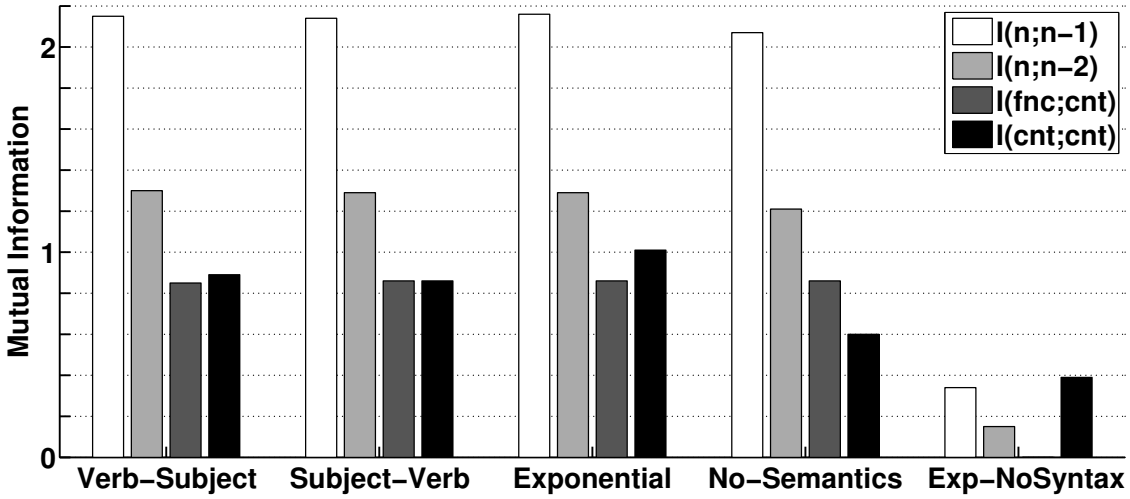


Figure 2-7: The mutual information words convey about other words in the same sentence. $I(n; n-1)$, in white, is the average mutual information which words in position $n-1$ of a sentence convey about their nearest neighbor in position n; the average is over all positions i within sentences. $I(n; n-2)$, in light gray, is the mutual information between words which are next nearest neighbors. The darker gray bars represent the amount of information $I(fnc; cnt)$ which function words (e.g. prepositions) convey about the nearest content word (e.g. a noun) in the same sentence. The black bars represent the amount of information $I(cnt; cnt)$ that content words convey about the nearest content word in the same sentence. For this last measure, note the considerable higher values for the semantics models. Note also that the mutual information between content words due to syntax and to semantics in the Exp model happens to be nearly equal to the sum of that due to syntax alone (in the No-Semantics model) and that due to semantics alone (in the Exp-NoSyntax model).

$I(n; n-1)$ is the mutual information which words in position $n-1$ of a sentence convey about their nearest neighbor (in position n); the average is over all positions in the sentence. As shown, this measure is almost the same for the semantics models (the absolute value of their difference is about 0.01), somewhat lower in the No-Semantics model (about 4% less), and considerably lower in the Exp-Nosyntax one (85% less), thus pointing at the strong effect of syntax. Likewise, we see a similar pattern across models for $I(n; n-2)$, which is the mutual information between words which are next nearest neighbors. The semantics models show again very similar statistics (with at most a 0.01 difference in absolute value), the No-Semantics is about 7% lower, and the Exp-Nosyntax model significantly decreases, by 88%.

We also show in Figure 2-7, the amount of information which function words (e.g.

prepositions) convey about the nearest content word (e.g. a noun) in the same sentence, $I(fnc, cnt)$. As shown, all the semantics models show the same $I(fnc; cnt)$ as the No-Semantics one, because there is no semantically induced component of the joint probability between function and content words, which reflects only syntax. In the Exp-Nosyntax model, this measure is zero, because in this particular model there is no syntax, while semantics only constrains pairs of content words.

An important result follows from considering the information that content words convey about the nearest content word in the same sentence, $I(cnt; cnt)$. We see considerably higher values for the semantic models than for the No-Semantics one. The Exponential model, in which all preceding words contribute, shows the heaviest dependence of word choice on previous history. More precisely, the Verb-Subject, Subject-verb, and the Exponential convey about 48%, 43%, and 68%, respectively, more information than the No-Semantics model. Thus, introducing semantics in the BLISS models further constrains the mutual dependence between content words by an additional 40-70%, without having other appreciable statistical effects. However, one should note that most of the mutual dependence between content words is already imposed by the syntax.

2.4.3 Memory Characteristics

To see how much our language models deviate from a first-order Markov model, we calculated the triple mutual information for 7-word sentences taking the Subject-Verb model as an example (Figure 2-8). The other models show very similar memory characteristics, the semantics models with very same numbers (the absolute value of the difference hovering around 0.01) and the No-Semantics with slightly less information (by about 0.09). In this analysis, the random variables are the words appearing in specific positions of a sentence. In Figure 2-8, each bar indicates the amount of entropy of the individual random variables, $H(n)$. In each bar, the black portion shows the amount of conditional mutual information, $I(n; n-2|n-1)$, whereas the gray portion shows the amount of mutual information, $I(n; n-1)$ (the sum of these two is just the triple mutual information $I(n; n-2, n-1)$). In a first order Markov sequence, the conditional information would be zero, so the non-zero

height of the black portions quantifies the non-Markovian character of BLISS.

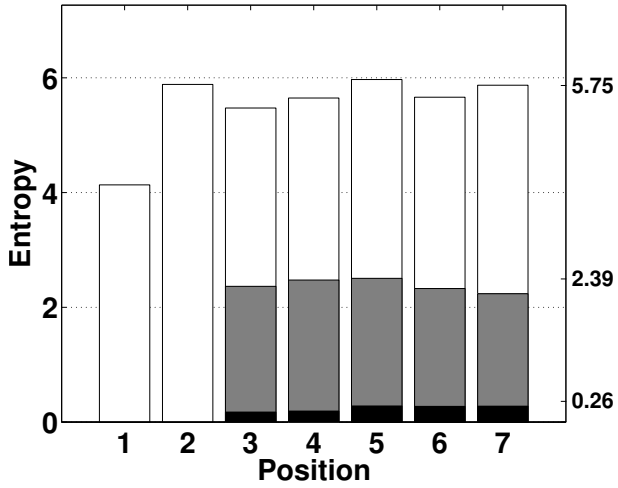


Figure 2-8: Markov properties of the Subject-Verb model. The full height of each bar indicates the entropy of individual words at each position in the sentence, $H(i)$. The gray portion of each bar shows the amount of mutual information, $I(i; i-1)$, whereas the black portion shows the amount of conditional mutual information, $I(i; i-2|i-1)$. Note that the sum of the black and gray portions equals the triple mutual information, $I(i; i-1, i-2)$. In a first order Markov sequence, the conditional information should be zero. The black portions of the bars indicate that BLISS generate sequences that deviate from first order Markov models. The numbers on the right vertical side of the box indicate the average of each measure over positions 2:7 for $H(n) = 5.75$, and over positions 3:7 for $I(n; n-1, n-2) = 2.39$ and $I(n; n-2|n-1) = 0.26$.

Figure 2-9 shows the dependence of the word at each position on those at each of the preceding positions, by calculating the pairwise mutual information between words at different positions within each sentence. Each shade represents the amount of information conveyed by words in a particular position about words in the following positions. For example, the (bottom) black portion of the bar in position i represents the information conveyed by words in position 1, namely $I(i; 1)$. Memory effects are seen to "decay" with position along the sentence; the effect is most noticeable for the subsequent 1-2 words, although an initial rapid drop in the degree of dependence is followed by less decay. The memory effect is relatively smaller for the first position compared to the information conveyed by the other position because the first position is usually occupied by the determiners (e.g. *the* or *a*), not having any semantic effect. Although these results are for the Subject-Verb model, similar results hold for the Verb-Subject and Exponential models (not shown).

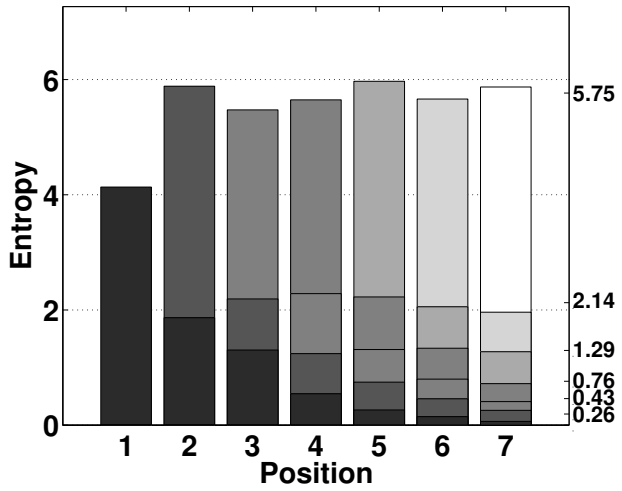


Figure 2-9: Memory effects in the Subject-Verb model. The dependence of each word on words at preceding positions is quantified by calculating pairwise mutual information values between positions along the sentence. Each shade represents the amount of information conveyed by words in a particular position about words in subsequent positions. For example, the black portion of the bar at position i, represents the information conveyed by words in position 1, namely $I(i;1)$. The numbers on the right of the box indicate the average values $I(n;n-i)$ over positions $i+2:7$: e.g. $H(n)=5.75$, $I(n;n-1)=2.14$, $I(n;n-2)=1.29$, and so on.

2.5 Discussion

We have constructed an artificial language, BLISS, which can be used in language learning experiments. BLISS is based on the putative universal principles of natural languages: vocabulary, syntax, and a definition of semantics, that is the statistical dependence between words in the same sentence. To make it as natural as possible, within its limited size, the probability of grammar rules, words, and word-pairs were extracted from two natural language corpora: the Wall Street Journal and Shakespeare.

The BLISS grammar is a Probabilistic Context-Free Grammar; in terms of the complexity, natural languages such as English have been argued to require (in order to be modelled) at least the level of complexity of context free grammars in the Chomsky hierarchy (Chomsky, 1957), though the grammars of some natural languages such as the Swiss German language are believed to be situated between context free and context sensitive because of their cross-serial clause construction and case-marking of objects (Shieber, 1985). The BLISS grammar also includes several phrasal and lexical categories which exist in natural

languages. There are both function words (e.g. prepositions, articles, ...) and content words (e.g. nouns, verbs, ...) in the grammar. The relation between function and content words have been of interest to cognitive neuroscientists (Hochmann et al., 2010). The grammar also contains a recursive construction, which might be used in experiments to investigate the effect of hierarchical structures in language processing (de Diego-Balaguer et al., 2011; Bahlmann et al., 2008).

Although the size of the BLISS grammar is quite larger than the size of toy grammars often used in artificial language learning studies, it is much smaller than the size of grammars developed in natural language parsers (e.g. the PCFG Stanford parser (Klein and Manning, 2003) or Charniak Parser (Charniak, 2000)). While obviously the grammar can easily be extended to have a larger number of production rules, in the current implementation of BLISS, we kept the size of the grammar as such to have a better control on the number of rules underlying the BLISS sentences and more importantly not to exceed the *storage capacity* of an associative neural network—the maximum number of words (and rules) that can possibly be stored in an artificial neural network (see section 3.2.1). These sentences will be used as the training samples for the neural network that we will introduce in the next chapter.

BLISS contains semantics which could be switched on and off, if the interaction between syntax and semantics is investigated in an experiment. Semantics is taken here to be the statistical dependence between words in the same sentence, produced by imposing additional, non-syntactic constraints on word choice during sentence generation. We have defined different models for semantics: Subject-Verb, Verb-Subject, and Exponential. In the Subject-Verb and Verb-Subject models, the importance of the thematic role was considered, as studied in several language learning experiments (Altmann and Kamide, 1999; Bicknell et al., 2010).

The semantics models defined in this study have been partially inspired by the well-known observation that highly predictable words in a context are easier to comprehend (Ehrlich and Rayner, 1981; Kutas and Hillyard, 1984). In order to better evaluate how cognitively plausible language models are, in general, one can compare their predicted reading time against the one needed by human participants in a reading task, for example, by using

surprisal theory (Hale, 2001; Levy, 2008). This theory predicts that the comprehension difficulty of word w_n in its sentential context $w_1 \dots w_{n-1}$ is proportional to the surprisal of the word, defined as $-logP(w_n|w_1 \dots w_{n-1})$; surprisal is minimized (goes to zero) when a word must appear in a given context (i.e. $P(w_n|w_1 \dots w_{n-1}) = 1$). The surprisal value can be estimated by the semantics models, which assign a selection probability to a word.

Our definition of *semantics* in this study is quite different from the common usage of this word in (computational) linguistics, where it refers to the "meaning" of a word, a sentence, or a discourse (a collection of sentences) (Jurafsky and Martin, 2000), concerning the formal representation and extraction of the knowledge that these linguistic entities convey about the world. In our definition of semantics, the co-occurrence of word pairs is of concern: the higher the co-occurrence of a word pair, the closer their semantic similarities. This approach is closer to machine learning approaches that disambiguate the sense (meaning) of a word with regard to the co-occurrence of the word with its neighbouring words in a discourse; note that, for having a better control and for simplicity, we consider only the words of the same sentence as neighbours. Further, the statistical dependence used in our study is different from the lexical and contextual information given to natural language parsers in order to augment their performance in assigning the most probable syntactic structure (e.g. a parse tree) to a sentence. In BLISS, after the grammatical structure of a sentence is fixed, we incorporate contextual information into the sentence under generation.

To measure the purely distributional effect of semantics, we have used information theory, which was extensively used in natural language processing studies, from document retrieval (Manning et al., 2008) to the evolutionary basis of natural languages (Piantadosi et al., 2011). We find that with this limited vocabulary, the effects of semantics are small relative to those produced by syntax.

We emphasize that in order to quantify the difference between models in terms of pairwise statistics (using either KL-divergence or mutual information), we need to have equal first-order statistics across the models. As discussed in the Appendix, this prevents us from setting the semantics coefficient at $g = 1$; in other words, we cannot use g to arbitrarily increase the effects of semantics and still be able to compare quantitatively with a language model without semantics, and have all models sound natural in terms of word frequency.

The only way we could make semantics more important in a linguistically plausible way is to increase the vocabulary beyond the ”intermediate” size which was our BLISS target.

Grammar and semantics concur in defining the memory properties of the language. Although they cannot be characterized as first-order Markov chains, the different variants of BLISS come relatively close to that, as suggested by the fact that the mutual information between two next-nearest neighbor words, conditional to the intermediate word, is very small, and that non-conditional mutual information between more distant words decays rapidly and then is followed by less decay. These results are shown in Figs. 2-8 and 2-9 for the Subject-Verb model, but they also apply to the other semantics models. It is an interesting question for future work to understand whether such rapid decay is characteristic of artificial languages modeled after English, and whether natural languages with more articulated syntax would lead to models with more persistent memory effects. We plan to address this question within the quantitative BLISS framework by introducing syntactic variants modeled after other natural languages.

Beside grammar and semantics, another main component of natural languages is of course phonology (the structure of sounds), which is not included in the current implementation of BLISS. Several artificial language learning studies (Monaghan et al., 2005; Toro et al., 2008; Shukla et al., 2011) investigate the importance of phonological cues in language acquisition. For instance, Monaghan et al. (Monaghan et al., 2007) investigate the interaction between phonological and distributional cues in learning lexical categories. A possible variant implementation of BLISS may include replacing real words by non-words which are controlled for phonological cues while maintaining distributional cues such as the joint probability of words.

Using parallel programming libraries and objected-oriented programming in python, BLISS software can generate millions of sentences in a few minutes. This provides a large sample of sentences which not only mimic natural languages but are also produced under the control of the experimenter.

BLISS can be used as a training sample for language acquisition by synthetic systems, including neural networks. Also, it can be used in experiments with adults when complex sentences are needed; in addition, with humans, its words can be replaced by pseudo-

words, or by visual or haptic signs. One may also modulate the complexity of a sentence, which is determined not only by the grammatical structure but also by its semantics, so as to study the processing of semantically reversible and irreversible sentences (Richardson et al., 2010).

An important feature of BLISS design, not elaborated here, is that it can be easily varied in its internal parametric structure. In its current version, BLISS is syntactically modeled after English, irrespective of the specific semantic model used. We can however keep the semantics stable, and model other natural languages by altering the production rules, either solely in the ordering of specific elements or also in their type and relative probability. Large corpora may then be produced that maintain a controlled and quasi-naturalistic flavor, while enabling the study of the relative learnability of distinct parameter settings (Gruening, 2004). Long-term, this may allow a scaled-down artificial language counterpart to the grand programme of studying the evolution of language diversity (Longobardi and Guardiano, 2009).

2.6 Conclusion

We have constructed BLISS, an artificial language of limited complexity that mimics natural languages by possessing a grammar of about 40 production rules, a vocabulary size of 150 words, and semantics defined by imposing dependencies between words. The effect of introducing semantics with such a limited vocabulary was quantified using methods of information theory and found to be small, but still measurable.

The code for generating BLISS is freely available from the authors. It is hoped that it will be used in a variety of ALL studies.

2.7 Appendix

2.7.1 Full Grammar of BLISS

All the grammar rules and the lexicon of BLISS are shown in Table 2.4.

2.7.2 Proof of Eq. 2.12

$$
\begin{aligned}
P_{post}(w_j) &= \sum_h P(h)P(w_j|h) \\
&= \sum_h P(h)\left((1-g)P_{prior}(w_j) + gP_s(w_j|h)\right) \\
&= \sum_{h_{ie}} P(h_{ie})P_{prior}(w_j) \\
&+ \sum_{h_e} P(h_e)\left((1-g)P_{prior}(w_j) + gP_s(w_j|h_e)\right) \\
&= \left(\sum_{h_{ie}} P(h_{ie}) + (1-g)\sum_{h_e} P(h_e)\right) P_{prior}(w_j) \\
&+ g\sum_{h_e} P(h_e)P_s(w_j|h_e)
\end{aligned}
\tag{2.13}
$$

After analyzing the relation between prior and posterior probabilities of a word in our models (Eq. 2.12), we have generated corpora with the desired overall word frequencies (the word frequencies of the semantics models were adjusted to the ones of the No-Semantics). Note, however, that there are some constraints regarding the possibility of generating sentences with arbitrary word frequencies. In Eq. 2.12, the numerator cannot be negative, because the output is a probability. Therefore, the possible posterior probability of a word is constrained by the parameter g and the pairwise statistics in $P_s(w_j|h_e)$, which here has been derived from Shakespeare.

2.7.3 Length of BLISS Sentences

The probability distribution of the length of sentences generated by the BLISS grammar is shown in Figure 2-10. To calculate mutual information values between words appearing in different positions of a sentence, we need to work with sentences of the same length. Note that the average sentence length is about 5. To obtain at least 5 distinct measures of triple information values, we decided to use sentences of at least 7 words for all calculations involving mutual information and triple mutual information.

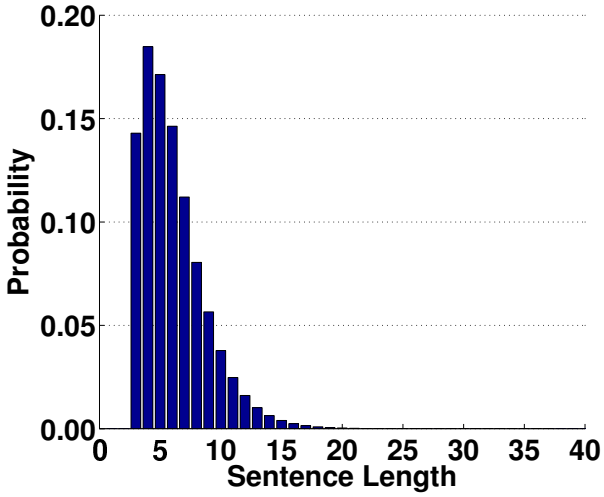


Figure 2-10: Probability distribution of sentence length in a corpus produced by the the Subject-Verb model. Other models with grammar show similar distributions.

2.7.4 Size of Model-Generated Corpora

A technical question to be addressed is how many sentences are needed to generate sufficient statistics. To answer this question, we produced up to 40 million sentences, and calculated several of the measures used in this paper.

In Figure 2-11, the mutual information between words in neighboring positions in a sentence, averaged over positions, $I(n; n-1)$, was measured for each model and with corpora of different sizes: 1, 10, 20, and 40 million sentences of at least 7 words. As shown below, 10 million sentences is enough for capturing pairwise statistics in the corpora, regardless of having syntax or semantics.

Figure 2-12 shows the average triple mutual information among words, $I(n; n-2, n-1)$. This quantity required the largest samples among all measures we have computed. As illustrated, increasing the size of the corpus from 20 to 40 million sentences does not appreciably change results for the models with syntax (Subject-Verb, Verb-Subject, Exponential, and No-Semantics), while the models without syntax are still changing and need even larger samples.

Corpora with 20 million sentences of at least 7 words each were used for the calculations of mutual information and triple mutual information in this study.

The KL-divergence calculations also require sufficient word-pair statistics. For those,

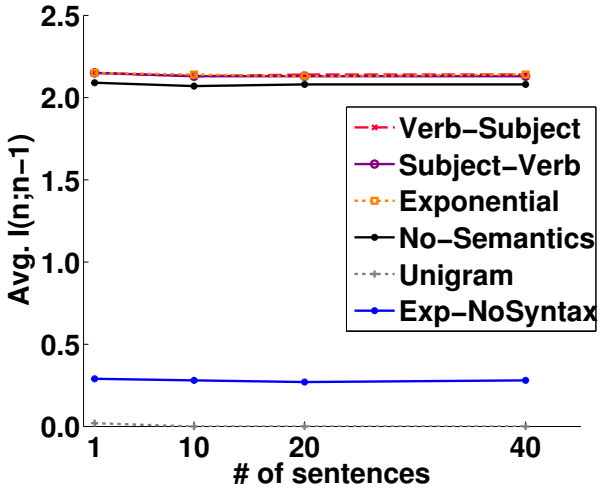


Figure 2-11: Average mutual information $I(n; n-1)$ averaged over all positions, versus number of sentences in the corpus. The mutual information between words is measured for each model and with corpora of different size: 1, 10, 20, and 40 million sentences of at least 7 words. As shown, 10 million sentences are enough for capturing pairwise statistics in the corpora, regardless of syntax or semantics. Note that the curves for the semantics models (Verb-Subject, Subject-Verb, and Exponential) are superimposed.

we have used corpora of at least 20 million sentences for the KL-divergence calculation as well, but without the 7-word constraint.

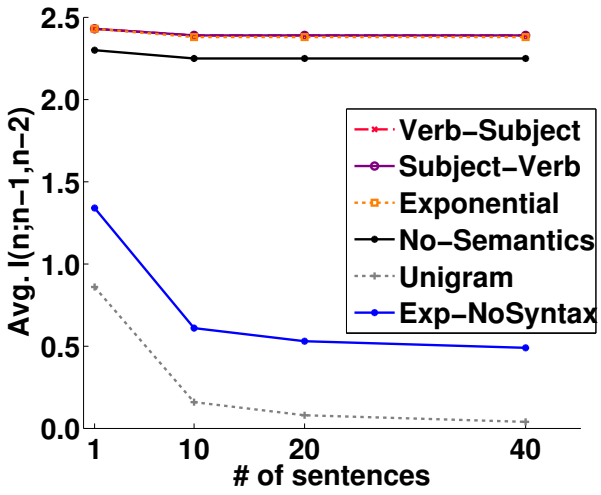


Figure 2-12: Average triple mutual information $I(n; n-1, n-2)$ averaged across positions, versus the number of sentences in the corpus. The three-way mutual information needs the largest sample among all calculated measures. As illustrated, increasing the size from 20 to 40 million sentences does not change it significantly for the models with syntax (Subject-Verb, Verb-Subject, Exponential, and No-Semantics). We see still some change for the models without syntax, which obviously need larger corpora, due to their larger word-to-word variability. Note that the curves for the semantics models (Verb-Subject, Subject-Verb, and Exponential) are superimposed.

Table 2.4: The full BLISS PCFG (Probabilistic Context-Free Grammar).

<S1>	→ <DP1><VP1>[0.50] | <DP+><VP+>[0.50]
<VP1>	→ <VR1>[0.85] | <Neg1><VR+>[0.15]
<VP+>	→ <VR+>[0.85] | <Neg+><VR+>[0.15]
<VR1>	→ <Vt1><DP>[0.37] | <Vtdtv1><DP><PPT>[0.06] | <Vtd1><SRd>[0.07] | <Vi1>[0.41] | <Vi1><PPV>[0.09]
<VR+>	→ <Vt+><DP>[0.36] | <Vtdtv+><DP><PPT>[0.09] | <Vtd+><SRd>[0.05] | <Vi+>[0.35] | <Vi+><PPV>[0.15]
<PP>	→ <Prep><DP>[1.00]
<PPV>	→ <PrepV><DP>[1.00]
<PPT>	→ <PrepT><DP>[1.00]
<SRd>	→ <Conjd><S1>[1.00]
<DP>	→ <DP1>[0.80] | <DP+>[0.20]
<DP1>	→ <Det1><NP1>[0.60] | <PropN1>[0.40]
<DP+>	→ <Det+><NP+>[0.20] | <NP+>[0.77] | <PropN+>[0.03]
<NP1>	→ <N1>[0.60] | <AdjP><N1>[0.20] | <N1><PP>[0.20]
<NP+>	→ <N+>[0.60] | <AdjP><N+>[0.20] | <N+><PP>[0.20]
<Det1>	→ <Art1>[0.97] | <Dem1>[0.03]
<Det+>	→ <Art+>[0.98] | <Dem+>[0.02]
<N1>	→ sword [0.13] | dagger [0.02] | crown [0.13] | dog [0.05] | horse [0.09] | dove [0.01] | calf [0.02] | deer [0.02] | worm [0.03] | sheep [0.03] | stone [0.04] | pearl [0.01] | wall [0.05] | gate [0.06] | house [0.18] | door [0.06] | rock [0.03] | church [0.04]
<N+>	→ swords [0.13] | daggers [0.02] | crowns [0.13] | dogs [0.05] | horses [0.09] | doves [0.01] | calves [0.02] | deer [0.02] | worms [0.03] | sheep [0.03] | stones [0.04] | pearls [0.01] | walls [0.05] | gates [0.06] | houses [0.18] | doors [0.06] | rocks [0.03] | churches [0.04]
<PropN1>	→ Zarathustra [0.25] | AhuraMazda [0.25] | Yasmin [0.25] | Ahriman [0.25]
<PropN+>	→ Magi [0.50] | Greeks [0.50]
<Vi1>	→ comes [0.33] | fights [0.07] | goes [0.21] | dies [0.07] | enters [0.08] | stands [0.14] | sits [0.07] | dares [0.03]
<Vi+>	→ come [0.33] | fight [0.07] | go [0.21] | die [0.07] | enter [0.08] | stand [0.14] | sit [0.07] | dare [0.03]
<Vt1>	→ praises [0.02] | loves [0.23] | follows [0.04] | cuts [0.02] | kills [0.04] | hates [0.02] | trusts [0.02] | serves [0.03] | holds [0.07] | hangs [0.03] | guards [0.02] | leaves [0.06] | needs [0.02] | wears [0.05] | marries [0.01] | gets [0.04] | finds [0.05] | takes [0.13] | banishes [0.01] | keeps [0.09]
<Vt+>	→ praise [0.02] | love [0.23] | follow [0.04] | cut [0.02] | kill [0.04] | hate [0.02] | trust [0.02] | serve [0.03] | hold [0.07] | hang [0.03] | guard [0.02] | leave [0.06] | need [0.02] | wear [0.05] | marry [0.01] | get [0.04] | find [0.05] | take [0.13] | banish [0.01] | keep [0.09]
<Vtd1>	→ wishes [0.12] | believes [0.03] | doubts [0.05] | hopes [0.12] | proves [0.12] | knows [0.32] | thinks [0.24]
<Vtd+>	→ wish [0.12] | believe [0.03] | doubt [0.05] | hope [0.12] | prove [0.12] | know [0.32] | think [0.24]
<Vtdtv1>	→ gives [0.80] | brings [0.20]
<Vtdtv+>	→ give [0.80] | bring [0.20]
<Conjd>	→ that [1.00]
<Prep>	→ of [0.60] | in [0.25] | with [0.07] | on [0.08]
<PrepT>	→ to [1.00]
<PrepV>	→ in [0.40] | with [0.30] | for [0.30]
<Neg1>	→ doesn't [1.00]
<Neg+>	→ don't [1.00]
<Art1>	→ the [0.70] | a [0.30]
<Art+>	→ the [1.00]
<AdjP>	→ great [0.16] | sweet [0.12] | noble[0.10] | poor [0.09] | long [0.08] | foul [0.04] | gentle [0.07] | bloody [0.04] | worthy [0.04] | strong [0.04] | holy [0.04] | heavy [0.03] | gracious [0.03] | royal [0.04] | mighty [0.02] | dangerous [0.02] | foolish [0.02] | precious [0.02]
<Dem1>	→ this [0.42] | that [0.58]
<Dem+>	→ these [0.60] | those [0.40]

3

Encoding BLISS Words into a Potts Attractor Network

"Zarathustra loves a holy church."

– BLISS

How are we humans able to produce sequences of words, namely sentences? We would like to approach this question by implementing a neural network that stores the words of BLISS, our artificial language of intermediate complexity, and produces sequences of these words in an orderly fashion. What network we used and how we encoded the words into this network are the questions discussed in this chapter.

We start with a short review of neuroscience findings on how words are represented in the brain. Next, we explain the characteristics of the network that we will use for representing the BLISS words. Finally, we discuss the algorithm of word representation and the detailed steps taken to encode the words into the network.

3.1 Evidence from the Brain for Word Representation

Is a cortical area of a few square centimetres the only locus of word production and comprehension (Broca, 1861; Wernicke, 1874), or, in contrast, do all areas equipotentially contribute to all cognitive operations including language (Deacon, 1989)? None of these

extremes, Donald Hebb suggested (Hebb, 1949). The Hebbian model—supported by electrophysiological studies (e.g. (Ahissar et al., 1992; Deacon, 1992; Vaadia et al., 1995))—is based on three assumptions: coactivated neurons become associated; this association can occur between adjacent or distant neurons; the associated neurons will develop into a functional unit—a *cell assembly*. The cortex can be considered as an associative memory, in which some areas are connected with each other.

Incorporating the Hebbian idea of cortical function into neural theories of language, several studies have considered cell assemblies as the neurobiological representation of words (Braitenberg and Schüz, 1991; Abeles, 1991; Pulvermüller, 1999); the representation of a word would be neither restricted to a small cortical area nor equally distributed all over the cortex, instead, it can be distributed over several relevant areas. Yet, the question remains: what are findings from the brain about the organization of words?

In psycholinguistics, there is a general agreement that retrieving a word from the brain provides access to three types of knowledge: the word's meaning (semantics or conceptual knowledge), its grammatical properties, and its sound structure (Jackendoff, 1999). In our study, we will focus on the representation of the semantic and syntactic properties of a word in the brain.

The semantic and syntactic representations of words are vastly believed to be encoded spatially separated in the brain, at least in part (for an overview (Shallice and Cooper, 2011; Shapiro and Caramazza, 2004; Pulvermüller, 1999; Petersson et al., 2010)). The comprehension and production of *content words*—with concrete and well imaginable meaning including nouns, verbs, and adjectives—remain intact in some (agrammatic) patients, who have lost their ability to use *function words*—of grammatical use, without a concrete meaning, including articles, auxiliary verbs, prepositions, and so on (Friederici and Schoenle, 1980; Friederici et al., 2000). The studies of patients with morphological deficits—the misuse of grammatical inflections such as the plural marker *-s* for nouns and the past tense marker *-ed* for verbs—demonstrate that these patients have exactly the same problems with real nouns and verbs as with meaningless non-sense words (used as nouns and verbs), thus indicating a distinction between semantic and syntactic knowledge in the brain (Shapiro and Caramazza, 2003; Shapiro et al., 2000).

Neuropsychological data of patients with semantic category-specific deficits clearly indicate that focal brain lesions selectively affect the semantic categories of words (demanding different sensory modalities): object words vs. action words (for an overview (Vigliocco et al., 2011)); living vs. non-living (Warrington and Shallice, 1984); animals vs. tools (Damasio et al., 1996). To explain semantic category-specific deficits, the developed theories fall into two broad groups (for an overview (Mahon and Caramazza, 2009)): the sensory/functional theory holds that the sensory/functional dimension provides the fundamental principle for the organization of the semantic system (Warrington and Shallice, 1984; Shallice, 1988); whereas, the domain-specific hypothesis—though also considering the sensory areas relevant—assumes that evolutionary pressures have resulted in specialized mechanisms for distinguishing categories (animates, inanimates, tools) (Caramazza and Shelton, 1998).

Evidence from neuropsychological studies of agrammatic patients—with syntactic deficits—indicate that the syntactic impairment is highly selective: following brain damage to the left cerebral hemisphere, patients lose their ability to inflect verbs for tense, but not for subject agreement; fail to form Wh questions, but not Yes/No questions in some languages; fail to produce subordinating conjunctions, but not coordinating conjunctions (Grodzinsky, 1984; Biran and Friedmann, 2011). To explain such selectivity, the impairment in agrammatic production has been characterized in terms of the inability of agrammatic speakers to access the high nodes in the hierarchy of the syntactic tree (Friedmann, 2001; Pollock, 1989). The grammatical constructs of tense inflection and Wh questions require the high nodes of the tree; whereas the constructs of subject-verb agreement inflection and Yes/No questions (in Hebrew and Arabic) are located in the lower part of the syntactic tree, thus correctly produced by agrammatic aphasics. Investigations of the nature of the tree-like hierarchical structure of language have shown that in the brain regions involved in syntactic computation, the fMRI BOLD response linearly increases with the size of the constituents of sentences (Pallier et al., 2011), whereas the human sentence processing in a reading task seems to be insensitive to hierarchical structure (Frank and Bod, 2011).

How could one account for these selective language deficits that the brain demonstrates, while having in mind the Hebbian theory of distributed representation of words? Feature

representation might be an answer: words are represented as a collection of features in a distributed fashion all over the cortex (McRae et al., 1997). The validity of a featural approach to word representation has been widely investigated by experimental, computational, and analytical studies. fMRI images show that the pattern of neural activation for a concept overlaps with the neural activation of the corresponding features (e.g. action and color features in (van Dam et al., 2012)). Computational studies have successfully predicted neural images of words by merely superimposing the neural images of a set of features (Mitchell et al., 2008).

A two-level multi-modular attractor neural network has been Proposed as a model of feature retrieval (O'Kane and Treves, 1992): a module that represented a patch of cortex is a local auto-associative memory network composed of a set of units that are connected through short-range synaptic connections; the module stores features through a Hebbian learning rule as local activity patterns. Once each module locally stores and retrieves features, a global auto-associative memory that models the cortex associates modules through long-range synaptic connections and stores concepts (or words) in a distributed multi-modal fashion in the network. This modular model is in accord with the proposal by (Braitenberg and Schüz, 1991): to a first-order approximation, the cortex can be considered as a two-level network of local and global auto-associative memory.

Inspired by this study, we have represented words as a set of features on a Potts attractor neural network, as discussed in section 3.2. We explain in section 3.3 how we have represented our words in the network, in a fashion loosely inspired by the brain.

3.2 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* (Amit, 1992; Hopfield, 1982).

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 using an attractor neural network, we followed insights offered by neuroscience: the power of attractor dynamics in turning analogue into nearly discrete operations to represent discrete concepts or words, the usage of the biologically plausible Hebbian learning rule for the storage of the concepts, the distributed representation of patterns reviewed in section 3.1, and the robustness allowed by analogue computation with distributed representations.

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 (Kanter, 1988; Treves, 2005), perhaps appropriate for modelling the language faculty and other high-level cognitive functions (Fig. 3-1). The Potts network is a simplified two-level, local and global associative memory network (Fulvi Mari and Treves, 1998; O'Kane and Treves, 1992), 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 (Braitenberg and Schüz, 1991)). 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 (Kropff and Treves, 2006). 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 Fig. 3-2, which shows the overlap between the actual network activation and the activation pattern that characterises the stored patterns as

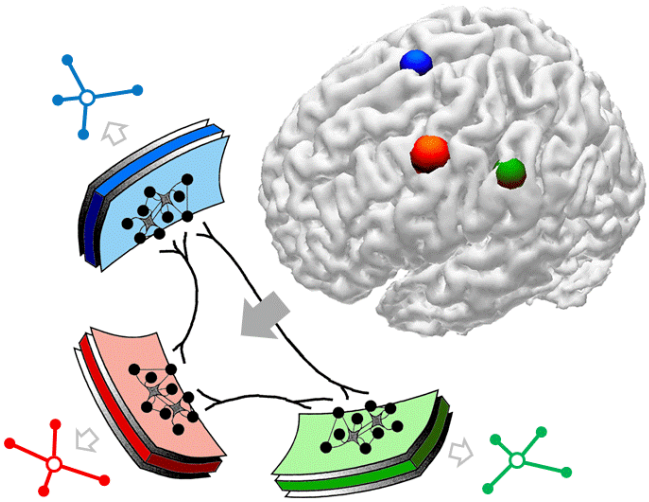


Figure 3-1: Conceptual derivation of the Potts network, which models the cortex as a simplified two-level, local and global associative memory network. The local dynamics of a cortical patch (coloured sheets) are reduced to a Potts unit (coloured units) with several states (here, 4). The global associative memory is the collection of the Potts units, which are connected through long range connections (black connections between cortical patches).

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.

In the following section we elaborate on the details of the network and the characterisation of its latching behaviour.

3.2.1 The Model

The Potts network is a collection of Potts units, which can be correlated to various degrees with any of S local attractor states—hypothetically, each state representing a local feature stored in a patch of cortex. The state variable of unit i, σ_i, is a $(S+1)$-dimensional vector: each of S elements of the vector measures how well the corresponding feature is being retrieved; the additional *zero-state* dimension provides the possibility of no correlation with any local attractor state (no significant retrieval). Since simultaneously a unit cannot

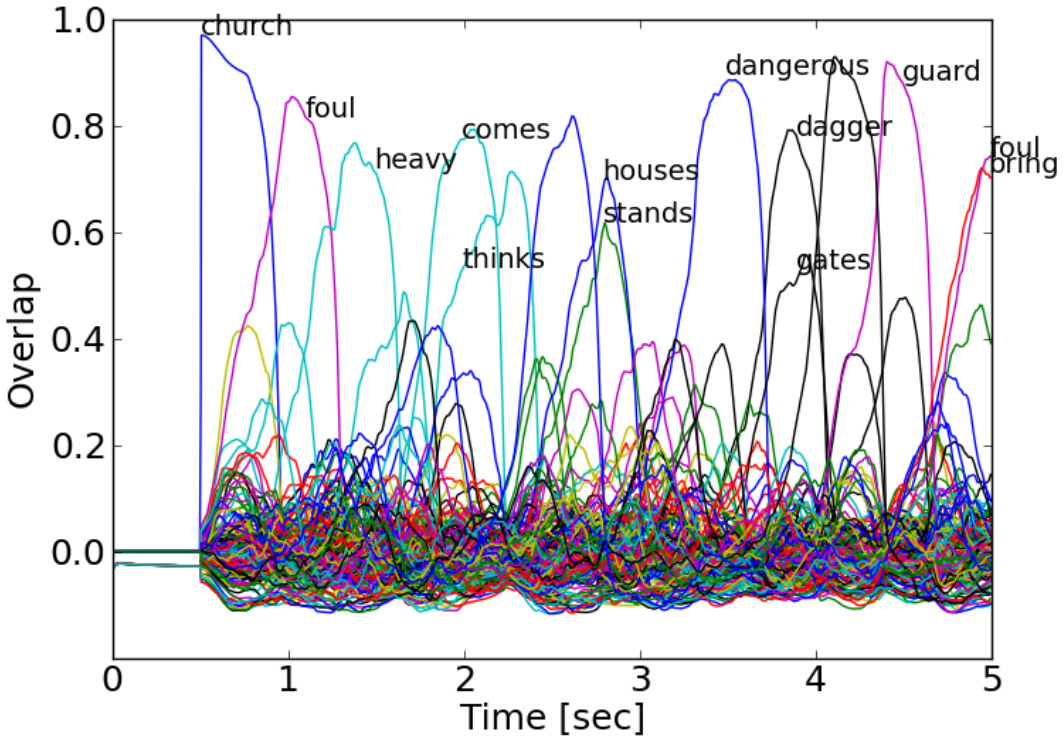


Figure 3-2: An example of latching dynamics in the Potts network. The x axis is time and the y axis is the correlation (overlap) of the network state with specific stored patterns, indicated in different colours. Each memory pattern represents a word. In this simulation the memory patterns are randomly correlated.

be fully active for all the states, we use a simple normalization $\sum_{k=0}^{S} \sigma_i^k = 1$.

Since features are represented as states of a Potts unit, the network stores a concept, or word, as a global attractor distributed over N Potts units. The representation of concept μ, where $\mu = 1, ..., p$, is indicated as $\xi^\mu := \xi_1^\mu, ..., \xi_i^\mu, ..., \xi_N^\mu$; each component represents the state of unit i—a discrete value which ranges from 0 to S—in concept μ.

Learning Rule

The network stores a concept through long-range synaptic connections between units, by associating the states of the units. For the storage of a concept, the strength of the connection between state k of unit i and state l of unit j, J_{ij}^{kl}—a quenched variable—is set to either associate a word to itself (features of a word to themselves), called *auto-association*, or associate a word to other words, called *hetero-association* (Sompolinsky and Kanter,

1986):

$$J_{ij}^{kl} = \frac{c_{ij}}{Ca(1-\frac{a}{S})}(c_{auto}\, Ja_{ij}^{kl} + c_{hetero}\, Jh_{ij}^{kl}), \tag{3.1}$$

where Ja_{ij}^{kl} is the auto-association component, with coefficient c_{auto}; Jh_{ij}^{kl} is the hetero-association component, with coefficient c_{hetero}; in the normalization factor, C is the average number of units connected to a given unit; a is the *sparsity* parameter—the fraction of units that are active in a pattern; c_{ij} is the element of the connection matrix—set to 1 if pre-synaptic unit j is connected to post-synaptic unit i, and 0 otherwise. The strength of the connection is normalized by the average number of *active* units connected to a given unit, Ca, and $1-\frac{a}{S}$ that is the maximum contribution by each of those units (see the next section).

Auto-associative learning rule. For the auto-association strength in the Potts network, each pattern μ is stored independently from other patterns with a contribution of $Ja_{ij}^{kl}(\mu)$ (Kropff and Treves, 2005):

$$Ja_{ij}^{kl} = \sum_{\mu=1}^{p} Ja_{ij}^{kl}(\mu), \tag{3.2}$$

where

$$Ja_{ij}^{kl}(\mu) = (\delta_{\xi_i^\mu k} - \frac{a}{S})(\delta_{\xi_j^\mu l} - \frac{a}{S})(1-\delta_{k0})(1-\delta_{l0}); \tag{3.3}$$

that is, the co-activation of two states k and l in two units i and j (of pattern μ) increases the synaptic strength between these two states. In Eq. 3.3, the Kronecker delta, $\delta_{\xi_i^\mu k}$, equals to 1 if the unit i of pattern μ is at state k, and it is 0 otherwise. The average activity of a state, $\frac{a}{S}$, is subtracted from the coupling constants to increase the *storage capacity*—the maximum number of stored and retrieved concepts—of an attractor neural network (Amit et al., 1987; Tsodyks and Feigelman, 1988; Fulvi Mari and Treves, 1998). The last two factors indicate that there is no synaptic strength from/to the null state ($k = 0$ or $l = 0$) (to avoid the activation of incoherent features, thus increasing the storage capacity (Kropff and Treves, 2005; Fulvi Mari and Treves, 1998)).

The associative learning rule that we use in the Potts network can be argued to have some biological motivation (Hebb, 1949), and it can be simply derived from the thermodynamical descriptions of the network. If the Hamiltonian (the energy function) of the

network, $H(\sigma)$, is defined as the weighted sum of the square of the overlaps, $m^{\mu}(\sigma)$, between network configuration σ and stored words μ, then by minimizing (and regularizing) the energy function of the network, the auto-associative strength is simply derived. For the Potts network, the energy function is (Kropff and Treves, 2005):

$$\begin{aligned} H(\sigma) &= -\frac{1}{2N}\sum_{\mu=1}^{p}(m^{\mu})^2 \\ &= -\frac{1}{2}\sum_{i,j\neq i}^{N}\sum_{k,l=1}^{S} J_{ij}^{kl}\sigma_i^k\sigma_j^l + U\sum_{i=1}^{N}\sum_{k=1}^{S}\sigma_i^k, \end{aligned} \tag{3.4}$$

where an overlap is defined as:

$$m^{\mu}(\sigma) \equiv \frac{1}{Na(1-\frac{a}{S})}\sum_{i=1}^{N}\sum_{k=1}^{S}(\delta_{\xi_i^{\mu}k} - \frac{a}{S})\sigma_i^k, \tag{3.5}$$

which varies between 0 (no correlation) and 1 (full retrieval).

Storage capacity. The storage capacity of a network is the maximum number p_c of words that can be stored in the network without significantly affecting the ability of the network to retrieve each pattern from a partial cue. If the number of stored patterns exceeds this limit, upon the presentation of a cue, the network retrieves a pattern that is uncorrelated or only partially correlated with the desired pattern.

For a fully connected Potts network, the capacity has been estimated to be $p_c \simeq NS^2$ (where S is the number of Potts states) (Kanter, 1988). Capacity estimations have been generalized to a variety of Potts networks by (Kropff and Treves, 2005) as a function of a (the sparsity) and S. In the *thermodynamic limit*, where N, p, $\beta \to \infty$, for $a \ll 1$, the storage capacity scales like $p_c \simeq CS^2/(4a)$, about $S^2/4$ times larger than the capacity of an auto-associative memory network with binary units (β is an inverse temperature that parametrizes the noise). More precisely, the capacity of a sparse Potts network is

$$p_c \simeq \frac{CS^2}{4\,a\,ln(\frac{2S}{a\sqrt{ln(S/a)}})}, \tag{3.6}$$

where C (the average number of connections per unit) can range between highly diluted connectivity ($C \simeq log(N)$) and full connectivity ($C = N - 1$). This estimation is validated by both numerical calculations and network simulations, though it only holds for randomly correlated patterns. For example, a Potts network with $C = 145$, $S = 7$, $a = 0.25$ can store around $7,000$ randomly correlated words.

In the case of correlated patterns, the basins of attraction of the correlated patterns overlap and it makes it difficult for the network to retrieve the stored patterns. The capacity of the Potts network for correlated patterns has been investigated in (Russo and Treves, 2012) through simulations: the presence of correlations (deviating from random correlations) significantly decreases the capacity of the network.

Hetero-associative learning rule. Beside associating each word to itself by using an auto-associative learning rule, we want to explore the possibility of associating words to each other using the hetero-association component Jh_{ij}^{kl} (Eq. 3.1). The hetero-associative learning rule has been used for storing a particular sequence of words (Sompolinsky and Kanter, 1986), and will be used here to incorporate *syntax* into the network. In the Potts network, the hetero-associative learning rule for storing only *one* sequence is:

$$\begin{aligned} Jh_{ij}^{kl} &= \sum_{\mu=1}^{p} Jh_{ij}^{kl}(\mu), \\ Jh_{ij}^{kl}(\mu) &= (\delta_{\xi_i^\nu k} - \frac{a}{S})(\delta_{\xi_j^\mu l} - \frac{a}{S})(1 - \delta_{k0})(1 - \delta_{l0}); \end{aligned} \tag{3.7}$$

note that feature ξ_j^μ of word μ is associated with feature ξ_i^ν of word ν, where μ precedes ν in the sequence.

The above hetero-associative rule for the storage of one particular sequence needs to be modified for the purpose of language modelling—specifically for storing more than one rule, as there are not one but many rules or valid sequences in natural languages. How many sequences can be stored in the network is a question that needs to be investigated separately. In section 3.3.4 we will explain how we modified this learning rule so as to partly incorporate syntax into the network.

Adaptive Dynamics

In the Potts network, a patch of cortex is modelled as a Potts unit. Though the internal neuronal dynamics of the patch is not incorporate in the model, a Potts unit is endowed with some realistic characteristics: adaptation and graded response.

Dynamic thresholds have been introduced in auto-associative networks to model neuronal fatigue and slow inhibition (Horn and Usher, 1989). In a Potts network (Treves, 2005), two types of dynamic thresholds are included for the activation of states in a Potts unit: (1) $\theta_i^k(t)$, which affects only active state k of unit i with a time constant of τ_2, intended to model resource depletion of the neurons and of the active synapses for that state (short-term depression (Tsodyks and Markram, 1997)); (2) $\theta_i^0(t)$, which affects the null state with a time constant of τ_3, intended to model slow inhibition within a cortical patch (Kohl and Paulsen, 2010). The dynamics of the time-varying thresholds are:

$$\tau_2 \frac{d\theta_i^k(t)}{dt} = \sigma_i^k(t) - \theta_i^k(t) \tag{3.8}$$

and

$$\tau_3 \frac{d\theta_i^0(t)}{dt} = \sum_{k=1}^{S} \sigma_i^k(t) - \theta_i^0(t). \tag{3.9}$$

The activity of active state k of unit i, $\sigma_i^k(t)$, is determined as,

$$\sigma_i^k(t) = \frac{exp(\beta r_i^k(t))}{\sum_{l=1}^{S} exp(\beta r_i^l(t)) + exp[\beta(\theta_i^0(t) + U)]}, \tag{3.10}$$

and for the null state as

$$\sigma_i^0(t) = \frac{exp[\beta(\theta_i^0(t) + U)]}{\sum_{l=1}^{S} exp(\beta r_i^l(t)) + exp[\beta(\theta_i^0(t) + U)]}. \tag{3.11}$$

Here $r_i^k(t)$ is the *firing rate*, U is a constant threshold that when added to $\theta_i^0(t)$ can be thought of as an input to the null state or as a common threshold to the other states, and β denotes an inverse temperature (T) that parametrizes the noise ($\beta = T^{-1}$). The activation functions are normalized such that $\sum_{k=0}^{S} \sigma_i^k(t) = 1$.

The firing rate of active state k of unit i, $r_i^k(t)$, rapidly integrates the *local field* $h_i^k(t)$,

the summed influence of pre-synaptic units, with a time constant τ_1, subject to its dynamic threshold $\theta_i^k(t)$:

$$\tau_1 \frac{dr_i^k(t)}{dt} = h_i^k(t) - \theta_i^k(t) - r_i^k(t), \tag{3.12}$$

where,

$$h_i^k(t) = \sum_{j \neq i}^{N} \sum_{l=1}^{S} J_{ij}^{kl} \sigma_j^l(t) + w \left(\sigma_i^k(t) - \frac{1}{S} \sum_{l=1}^{S} \sigma_i^l(t) \right). \tag{3.13}$$

The first term in Eq. 3.13 accumulates the inputs to the patch represented as unit i, the second term or the *local feedback term*, with coefficient w, provides a positive feedback that models the non-linear convergence towards the more active state.

Latching phases

An auto-associative memory network with dynamic thresholds produces sequences of memory patterns (latching) though with a limited storage capacity (Herrmann et al., 1993). How can this latching ability be enhanced in a Potts network with much higher capacity? Not only can latching be composed of hundreds of patterns, (Treves, 2005) suggests that it can even continue *indefinitely*. This suggestion has been further considered in studies that investigated the complexity of latching transitions (Kropff and Treves, 2006; Russo et al., 2008) and the phases of latching length (Russo and Treves, 2012; Abdollah-nia et al., 2012).

The phase diagram of latching dynamics has been explored by (Russo and Treves, 2012) through both numerical simulations and computational analysis (Fig. 3-3). For small w, the Potts network only retrieves a cued pattern and does not latch to other patterns—*no latching phase*. For larger w, the network produces a finite number of transitions (*finite latching*), while a further increase in w results in an infinite number of transitions in the network (*infinite latching phase*). Exceeding another threshold, the network gets trapped in the retrieved pattern (*stable attractor phase*) (Hopfield, 1982).

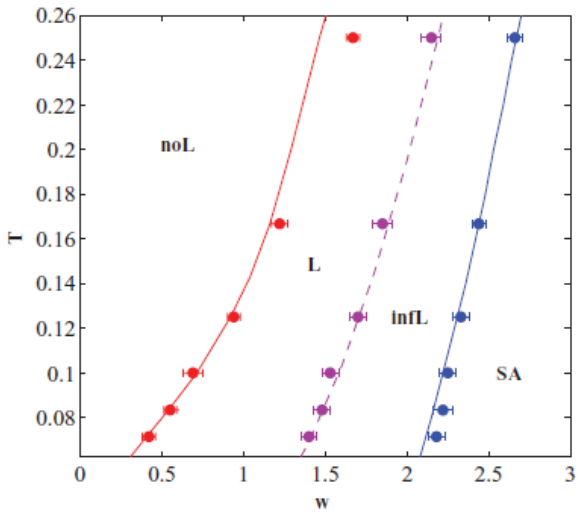


Figure 3-3: The latching phase diagram in a $w - T$ space: (from left to right) no Latching (noL), finite latching (L), infinite latching (infL), and stable attractors (SA) (Russo and Treves, 2012). Solid curves are analytical results, while the dotted curve is just a guide for the eye. Dots are the results of numerical simulations with parameters $S = 7$, $N = 600$, $p = 140$, $C = 90$, $a = 0.25$, and $U = 0.1$. A rough estimation of the storage capacity given these parameters is about 1200 memory patterns.

3.3 Implementation of Word Representation in the Potts Network

Having constructed the training language BLISS and implemented 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, 541 out of which express the semantic content, $N_{sem} = 541$ (comprising the *semantic sub-network*), and the rest, 359 units, represent the syntactic characteristics of a word, $N_{syn} = 359$ (comprising the *syntactic sub-network*). The distinction between the semantic and syntactic characteristics of a word has been loosely inspired by neuropsychological studies discussed in section 3.1. 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). While the sparsity (the fraction of active Potts units) was kept the same for all 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 ($a_{sem}^{fwd} = 0.08$, 45 active units out of 541 units) compared to the content words ($a_{sem}^{cnt} = 0.25$, 135 out of 541), whereas syntactic units are more active for the function words ($a_{syn}^{fwd} = 0.50$, 180 out of 359) than for the content words ($a_{syn}^{cnt} = 0.25$, 90 out of 359).

3.3.1 Generating Algorithm

In order to represent words in the network, we need a generating algorithm that reflects the variable degree of correlation between words. We used an algorithm comprised of two steps (Treves, 2005): (1) we establish a number of vectors called *factors* or features. Each factor influencing 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 the *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 question of how many hidden factors we needed—how uncorrelated the words needed to be—is related to the quality of the latching behaviour in a Potts attractor neural network. For a single latching transition, correlation between two patterns is necessary, as the transition is less likely between two orthogonal patterns (Kropff and Treves, 2006; Russo et al., 2008). Though the phase diagram of latching behaviour with correlated pat-

terns needs to be extensively investigated by later studies, our simulations indicate, that we can better control latching dynamics by keeping patterns close to the phase of random correlation.

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 (Mitchell et al., 2008), 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.

Evaluation of Word Correlations

Generating words using the above algorithm, we quantified the correlation between two words μ and ν, with N units, as

$$Nas^{\mu\nu} = \left\langle \sum_{i=1}^{N} \delta_{\xi_i^\mu \xi_i^\nu} (1 - \delta_{\xi_i^\nu 0}) \right\rangle_{\mu \neq \nu}, \tag{3.14}$$

the number of active units that are at the same state (*Nas*) in both patterns. We use the notation $< Nas >=< Nas^{\mu\nu} >_{\mu\nu}$ to measure the average correlation across all words of two word categories—either the same or different.

If two patterns are randomly correlated, we expect the correlation measure to be:

$$Nas^{\mu\nu} \simeq N \frac{a^\mu}{S} \frac{a^\nu}{S} S = N a^\mu \frac{a^\nu}{S}, \tag{3.15}$$

where a^μ is the sparsity of pattern μ and S is the number of active states that a unit of this pattern may occupy. Thus $\frac{a^\mu}{S}$ is the probability of a particular active state in a unit of pattern μ.

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 $a = 0.50$.

We now use the generating-algorithm to represent the 149 words in BLISS in the semantic and the syntactic Potts sub-networks.

3.3.2 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 (Fig. 3-4). For each word, 541 semantic units were used ($N_{sem} = 541$). Out of 541, 135 units became activated in all the content words ($a^{cnt}_{sem} = 0.25$); for the function words 45 units were activated ($a^{fwd}_{sem} = 0.08$).

For the representation of nouns we used the feature norms in the McRae database (McRae et al., 2005). 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

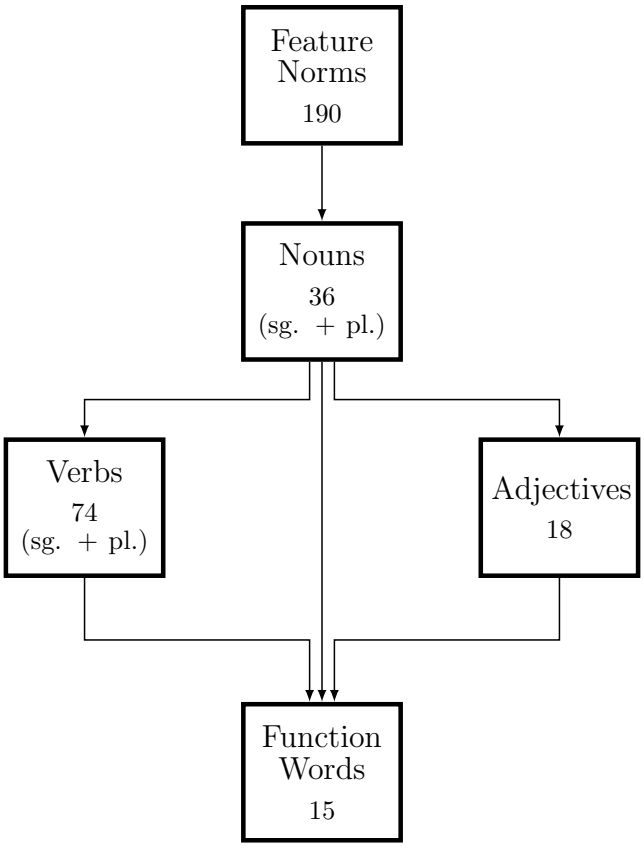


Figure 3-4: Order of word generation in the semantic sub-network. Beside the main factors, additional 400 hidden factors were used. The number of words in a word category is written below its name (for nouns and verbs, singular and plural forms were counted together). Proper nouns are not shown in the diagram because of their small size. All factors were generated with the sparsity $a = 0.25$, except for function words, $a = 0.08$.

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 fea-

tures 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.

3.3.3 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 (Fig. 3-5). For each word, 359 syntactic units were dedicated; out of 359, 180 units became activated in function words ($a_{syn}^{fwd} = 0.5$), whereas,

90 units were activated for the content words ($a_{syn}^{cnt} = 0.25$).

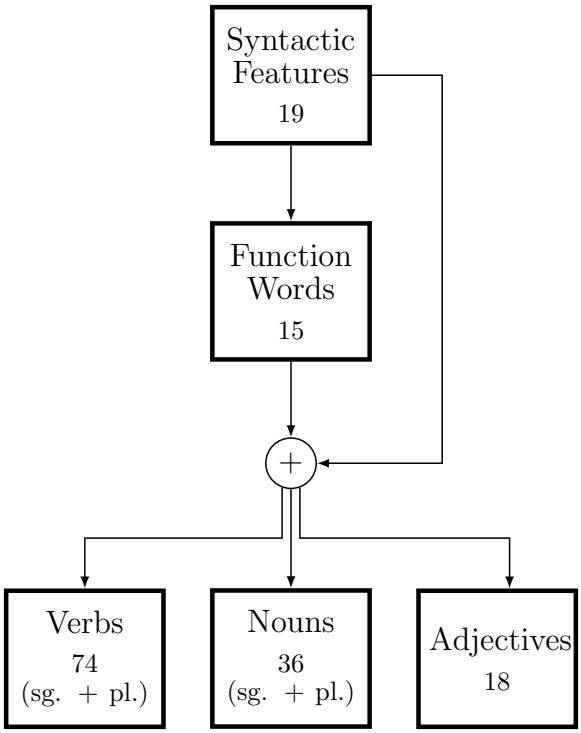


Figure 3-5: Order of word generation in the syntactic sub-network. Beside the main factors, additional 20 hidden factors were used. The number of words in a word category is written below its name (for nouns and verbs, singular and plural forms were counted together). Proper nouns are not shown in the diagram because of their small number. All factors were generated with the sparsity $a = 0.25$, except for function words, $a = 0.50$.

As factors for generating the function words, we arbitrarily designed 19 syntactic features:

- 7 lexical categories: lxc/noun, lxc/verb, lxc/adj, lxc/conjunction, lxc/preposition, lxc/pronoun, lxc/adverb
- 2 numbers: Number/singular, Number/plural
- 1 negation: Negation
- 3 determiners: Determiner/indefinite, Determiner/definite, Determiner/properNoun
- 2 locations: Location/close, Location/far
- 4 directions: Direction/from, Direction/towards, Direction/samePlace, Direction/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 3.1 in Appendix).

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.

3.3.4 The Potts Network stores the BLISS Words

For storing the BLISS words into the Potts network, we used the learning rule of Eq. 3.1, with both auto-associative and (wherever needed) hetero-associative components, and parameters $a = 0.25$, $S = 7$, and $p = 149$. The number of connections per unit (C) was kept around $C/N \approx 0.16$, in the range that is appropriate for the latching transition (Russo and Treves, 2012); we used $C_{sem} = 88$ for the semantic sub-network and $C_{syn} = 58$ for the syntactic sub-network, if no interaction between these two parts exists. If there an

interaction, these parameters changes (it will be discussed later).

Semantic sub-network

To store the semantic representation of the words into the Potts network, we used the auto-associative (Hebbian) learning rule in Eq. 3.2 and switched off the hetero-associative component ($c_{auto} = 1$ and $c_{hetero} = 0$ in Eq. 3.1). The only correction that we made to the rule in Eq. 3.2 was for the storage of the nouns and verbs: as the plural and singular forms of the nouns and verbs are identical in the semantic representation, we halved the contribution of a noun or verb to the synaptic strength to avoid double synaptic strengths for these words.

Syntactic sub-network

For the syntactic representation, we used both the auto-associative and hetero-associative components, setting $c_{auto} = 1$ and varying c_{hetero}. We added a normalization term to Eq. 3.1 because of high correlations between words belonging to the same syntactic category. As we will shown in the Results (Fig. 3-12), the average correlation of the syntactic component of words that belong to the same category is very high compared to that of randomly correlated patterns; for instance, the average correlation between words belonging to the singular noun category is about $< Nas >_{nsg} = 50$, while for randomly correlated patterns it is estimated to be $< Nas >_{nsg} = 3.2$ (see Eq. 3.15).

Because of such high correlation between the words that belong to the same syntactic category, the contribution of word μ (e.g. *sword*) that belongs to syntactic category γ (e.g. *nsg*) was normalized by the number of words in that category, k_γ (e.g. 18 for the singular nouns). Since the overlap between words in the same syntactic category is not full, we took the average correlation into account, to obtain

$$J_{ij}^{kl}(\mu) \Leftarrow \frac{1}{k_\gamma \, nas_\gamma} J_{ij}^{kl}(\mu), \tag{3.16}$$

where $nas_\gamma = \frac{<Nas>_\gamma}{maxNas}$ is the average correlation between the words that belong to category γ, normalized by the maximum Nas in the syntactic network (e.g. 90 for the singular nouns).

Beside auto-associating the words in the syntactic part (Eq. 3.2), we used the hetero-associative learning rule (Eq. 3.7) for storing the sequences of the words that appear consecutively in a BLISS corpus. In the corpus, if word μ precedes word ν with probability $P(\mu,\nu)$ ($\sum_{\nu=1}^{p=149} P(\mu,\nu) = 1$) then the contribution of word μ to the hetero-associative rule (Eq. 3.7) is modified as:

$$Jh_{ij}^{kl}(\mu) = (\delta_{\xi_j^\mu l} - \frac{a}{S})(1-\delta_{k0})(1-\delta_{l0}) \sum_{\nu=1}^{p} P(\mu,\nu)(\delta_{\xi_i^\nu k} - \frac{a}{S}). \tag{3.17}$$

Connection between Semantic and Syntactic sub-networks

Having specified the learning rule used for both semantic and syntactic sub-networks, we then connected these two sub-networks using the directional auto-associative learning rule (symbolized in Fig. 3-6).

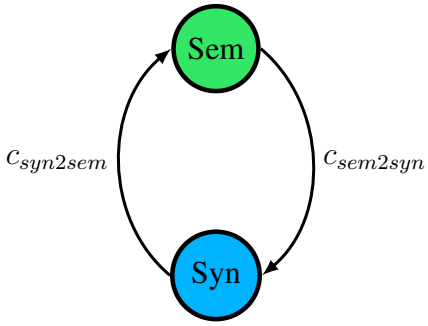


Figure 3-6: Interaction between semantic and syntactic sub-networks.

We have varied the strength of the connection from the semantic sub-network to the syntactic sub-network ($J_{i\in syn,j\in sem}$) by modulating the coefficient $c_{sem2syn}$, and the strength of the connection from the syntactic sub-network to the semantic one ($J_{i\in sem,j\in syn}$) by modulating $c_{syn2sem}$:

$$J_{i\in syn,j\in sem}^{kl} \Leftarrow c_{sem2syn} * J_{i\in syn,j\in sem}^{kl} \tag{3.18}$$

$$J_{i\in sem,j\in syn}^{kl} \Leftarrow c_{syn2sem} * J_{i\in sem,j\in syn}^{kl}. \tag{3.19}$$

Since this interaction modifies the number of connections per unit (C) in the two sub-networks, the number of connections for the Potts units of the semantic sub-network, C_{sem},

and for the units of the syntactic sub-network, C_{sem}, increase to:

$$C_{sem} \Leftarrow C_{sem} + c_{syn2sem} * C_{syn} \tag{3.20}$$

$$C_{syn} \Leftarrow C_{syn} + c_{sem2syn} * C_{sem}. \tag{3.21}$$

Without interaction ($c_{sem2syn} = 0.0$ and $c_{syn2sem} = 0.0$), their value are $C_{sem} = 88$ and $C_{syn} = 58$.

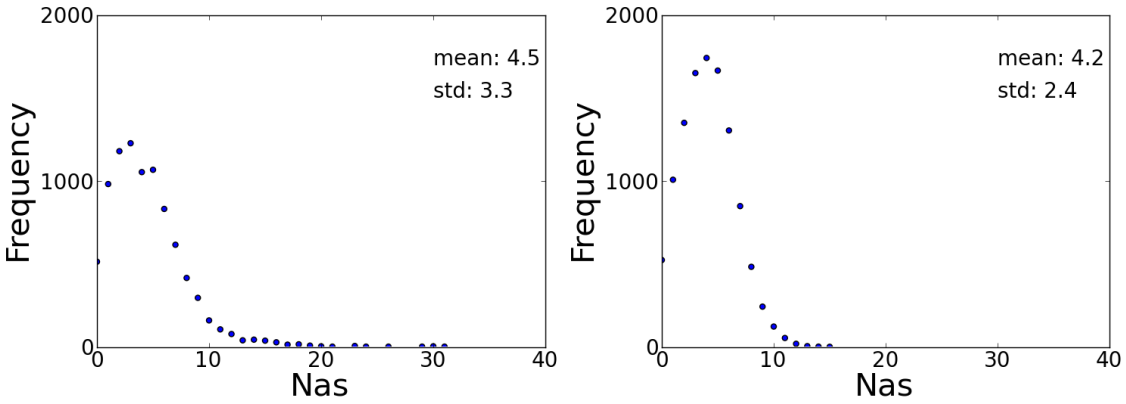
3.4 Results

After representing both semantic and syntactic components of the BLISS words, we measured the correlations between words first for the semantic component, then for the syntactic part.

3.4.1 Word Correlations in the Semantic Representation

After generating the semantic representation of words following the procedure discussed above, we compared (Fig. 3-7) correlations of the words with correlations of 149 randomly correlated patterns (having 134 patterns with $a = 0.25$ and 15 patterns with $a = 0.08$). These random patterns were generated by the algorithm suggested by (Treves, 2005), with parameters set so as to produce randomly correlated patterns with an average correlation close to the theoretical measure, $<Nas> = 4.1$ (Eq. 3.15). Although the average correlation of the words in the semantic sub-network of 4.5 is close to the expected correlation of random patterns, the distribution of correlations exhibits a tail with very high values. When computing the correlation distributions shown in Fig. 3-7, all 149 random patterns were used in Fig. 3-7b, but only the singular form of nouns and verbs in Fig. 3-7a. Plural forms were not included because they are semantically identical to their singular form; thus the discrepancy in frequencies in the two plots shown in Fig. 3-7.

To have a better understanding about the word correlations in the semantic representation, we first measured the correlation between nouns belonging to different semantic categories (*animals*, *buildings*, *objects*). Fig 3-8 demonstrates higher correlation of words



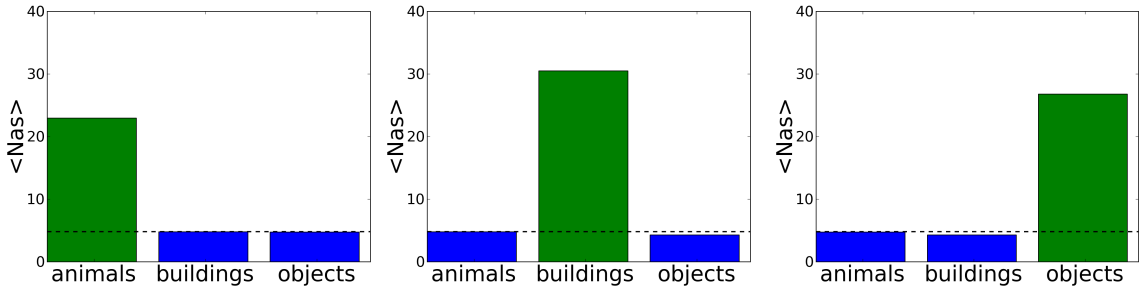
(a) The words in the semantic sub-network

(b) Randomly correlated patterns

Figure 3-7: Frequency distributions of correlations (*Nas*) (a) between the words in the semantic sub-network and (b) between randomly correlated patterns.

within a semantic category (in green) than across categories (in blue); the average correlations across the categories are close to the correlation of randomly correlated patterns, indicated by a dashed line. For instance, Fig. 3-8a compares the correlation of words belonging to the *animals* category (*calf*, *deer*, *dog*, *dove*, *horse*, *sheep*, and *worm*) with the words of either the same category (animals) or other categories of *buildings* (*church*, *door*, *gate*, *house*, and *wall*) and *objects* (*crown*, *dagger*, *pearl*, *rock*, *stone*, and *sword*). Note that the maximum correlation between every two content words—the maximum number of units that are active and at the same state in a pair of words—is 135 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 every factor with the generated word in the Potts network and in the training BLISS corpus, generated by the Subject-Verb model (Fig. 3-9). 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. 3-9 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

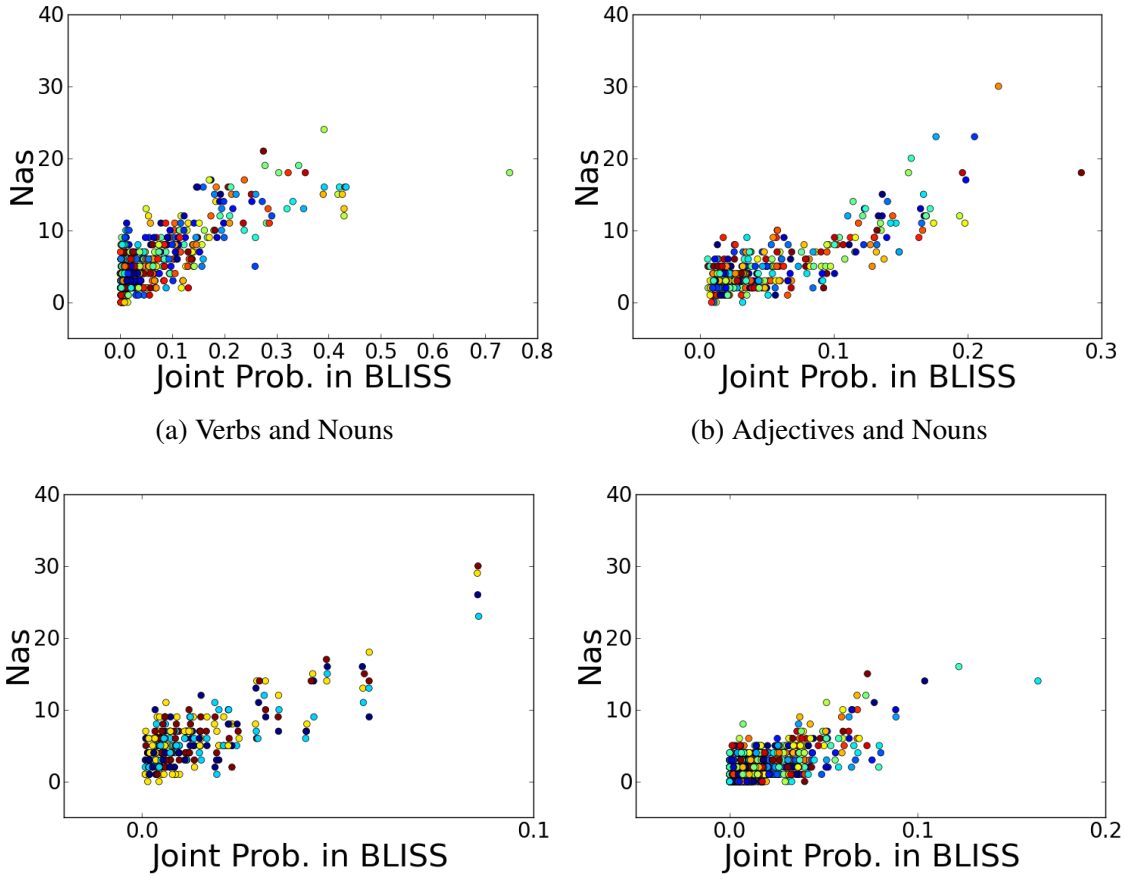


(a) animals (b) buildings (c) objects

Figure 3-8: The average correlation (<*Nas*>) between words within a semantic noun category (*animals*, *buildings*, and *objects*), shown in green, or across the categories, in blue. (a) The average correlation between the words of "animal" category (e.g. *dog* and *horse*) with themselves and with the words of "buildings" (e.g. *buildings* and *house*), and the words of "objects" (e.g. *sword*, *stone*); likewise, for (b) buildings and (c) objects. The dashed line indicates the expected correlation of randomly correlated patterns with $a = 0.25$ ($\simeq 4.8$).

of randomly correlated patterns ($<Nas> \simeq 4.8$, between content words; $<Nas> \simeq 1.5$, between content words and function words; Eq. 3.15).

After generating the semantic representation of all the words, we compared the average correlations (<*Nas*>) of the semantic representation of the words that belong to different word categories (nouns, verbs, adjectives, proper nouns, and function words) across the categories. As shown in Fig. 3-10, correlations between word categories are close to the regime of randomly correlated patterns in the semantic sub-network, that is $<Nas> \simeq 4.8$ between content words, and $<Nas> \simeq 1.5$ between content words and function words (see Eq. 3.15). However, the average correlation between words within a category is higher than the correlations across categories; this increase is more noticeable for proper nouns because of their small size (4 singular and 2 plural), and it is less noticeable for function words because of their small sparsity ($a_{sem}^{fwd} = 0.08$). The high correlations within categories indicate deviations of the generated patterns in the semantic sub-network from the randomly correlated regime.



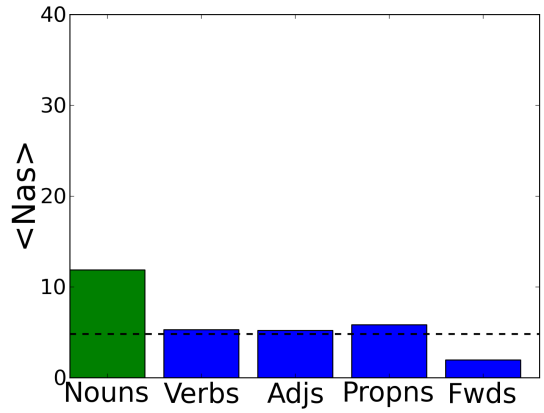
(a) Verbs and Nouns

(b) Adjectives and Nouns

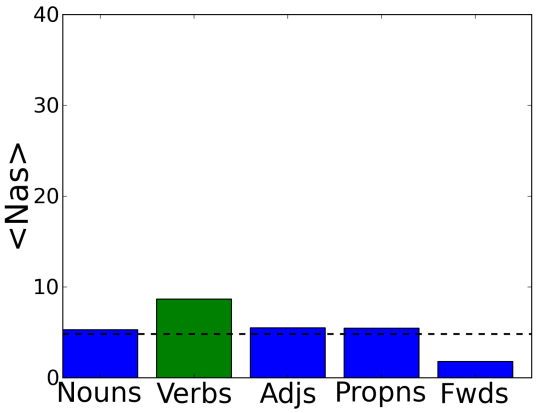
(c) Proper Nouns and content words

(d) Function words and content words

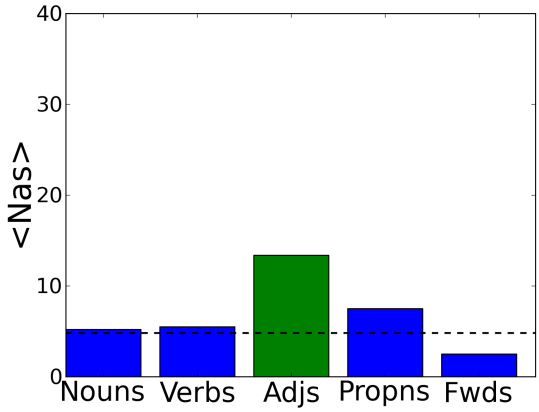
Figure 3-9: Comparison of the correlation between words and their factors in their semantic representation (<*Nas*> on y-axis) versus their joint probabilities in the BLISS corpus produced by the Subject-Verb model (x-axis) (the joint probabilities between a word and its factors have been normalized to 1). Each dot indicates a pair of a word, that is associated with a color, 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*), (c) the 4 singular proper nouns and other content words (nouns, verbs, and adjectives), which were the proper nouns' factors (e.g. *Zarathustra sword* or *Zarathustra kill*), and (d) the function words and their factors (nouns, verbs, and adjectives) (e.g. *the sword*).



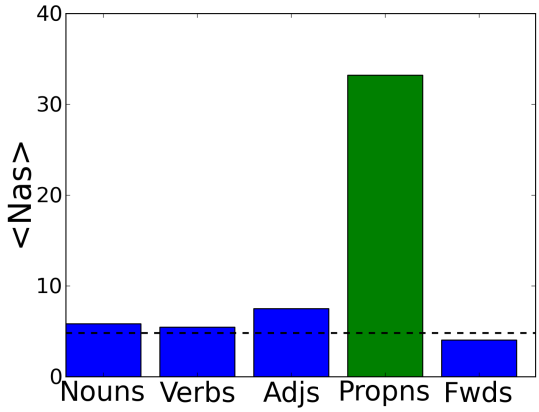
(a) Nouns



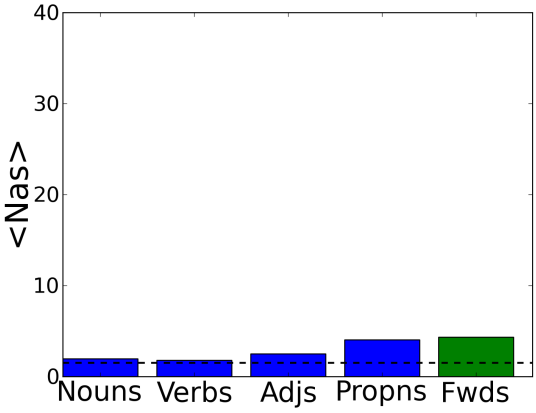
(b) Verbs



(c) Adjectives



(d) Proper nouns



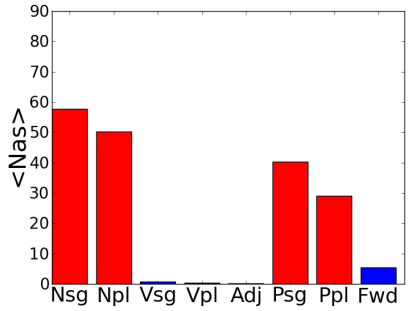
(e) Function words

Figure 3-10: The average correlation (<*Nas*>) between words within a category (in green) or across categories (in blue) in the semantic representation. (a) The average correlation of words that belong to the noun category with either themselves (nouns) or with words of other word categories (verbs, adjectives, proper nouns, and function words); likewise for other categories in (b)–(e). The dashed lines indicate the expected correlation of randomly correlated patterns ($\simeq 4.8$, between content words; $\simeq 1.5$, between content words and function words).

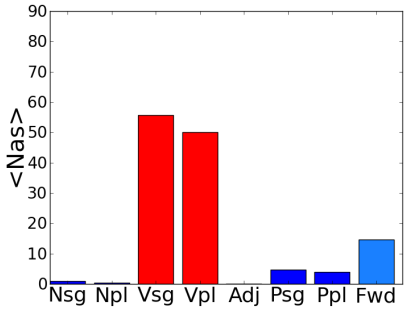
3.4.2 Word Correlations in the Syntactic Representation

For the syntactic representation of the words, we used 19 syntactic features and 20 hidden factors as factors. Among them, as shown in Table 3.1, the 7 *LexicalCategory* features have the strongest weights (lxc/noun, lxc/verb, lxc/adj, lxc/conjunction, lxc/preposition, lxc/pronoun, lxc/adverb). Fig. 3-11 demonstrates the average correlation, $<Nas>$, of each of these factors with words of different word categories. The colour code indicates the strength of the suggestion weights of factors for the syntactic categories, varying from the weakest (dark blue) to normal (green) and to the strongest (red). Comparing these correlations with the weights with which these factors influenced the words in Table 3.1, we observe that stronger weights naturally result in higher correlation between representations; for example, in Fig. 3-11a, *lxc/noun* is highly correlated with noun categories (singular and plural nouns and proper nouns). In the next chapter, we will take advantage of these high correlations for demonstrating the behaviour of the syntactic sub-network.

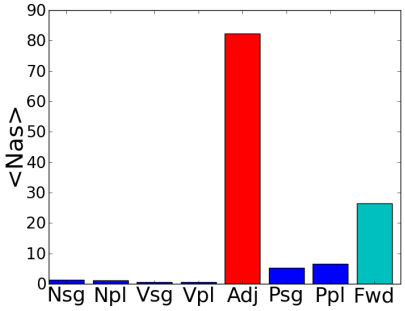
Generating the syntactic representation of all the words, we measured their correlations within and across different syntactic categories (singular and plural nouns, singular and plural verbs, adjectives, singular and plural proper nouns, and function words), as shown in Fig. 3-12. As expected, the correlations 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, see Eq. 3.15. As shown in Fig. 3-12a, 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.



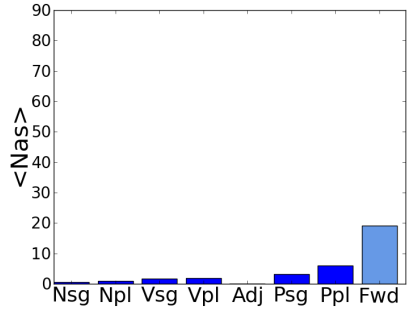
(a) lxc/noun



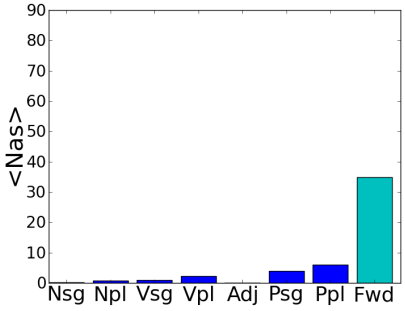
(b) lxc/verb



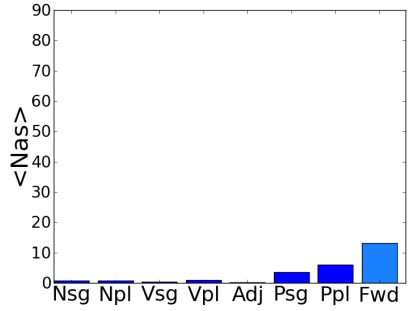
(c) lxc/adj



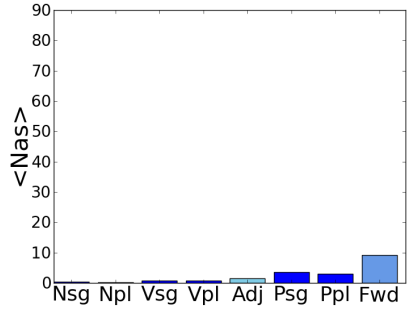
(d) lxc/conjunction



(e) lxc/preposition

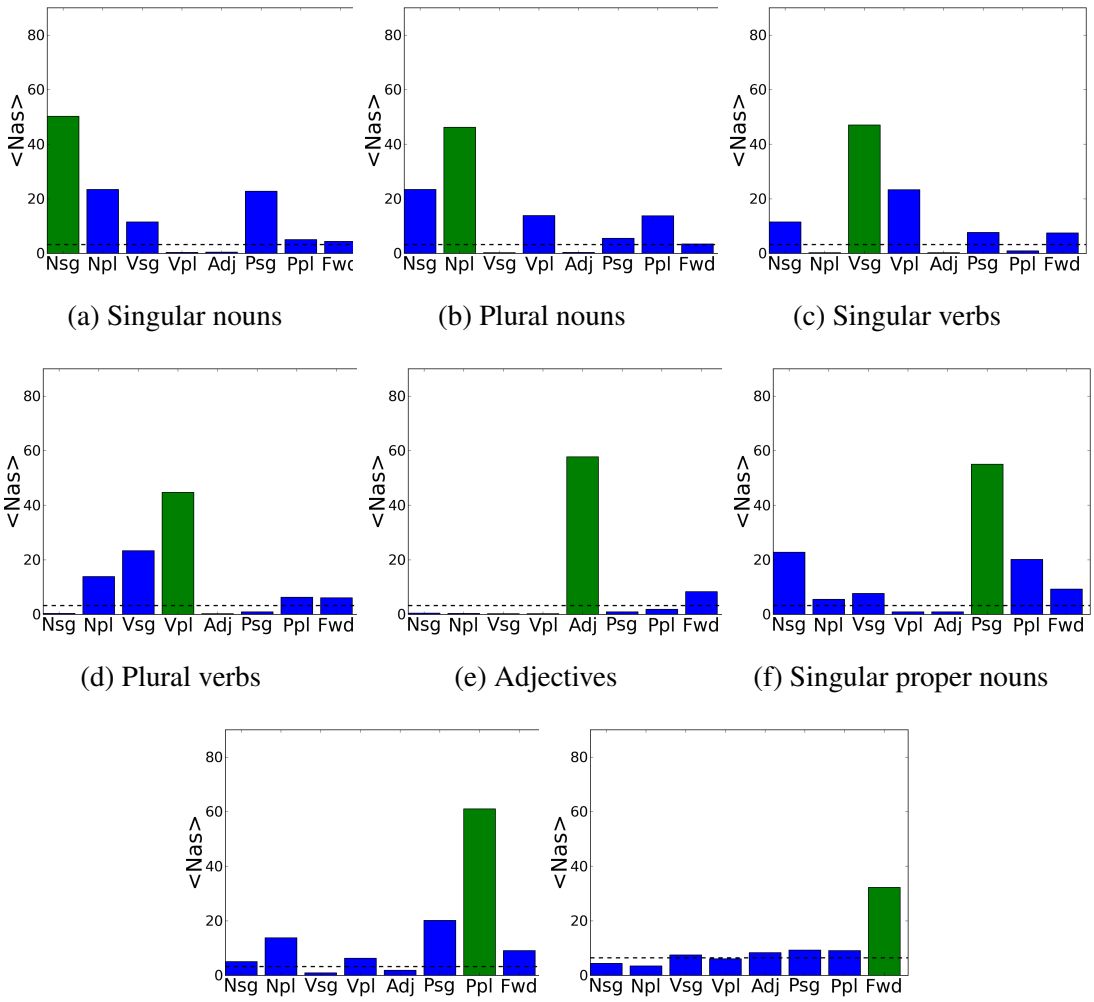


(f) lxc/pronoun



(g) lxc/adverb

Figure 3-11: The average correlation (<*Nas*>) of the 7 lexical categories used for the generation of the syntactic representation of words, with words belonging to different syntactic categories. The colour code indicates the strength of the suggestion weights of these factors for the syntactic categories (Table 3.1), varying from the weakest (dark blue) to normal (green) and to the strongest (red).



(a) Singular nouns (b) Plural nouns (c) Singular verbs

(d) Plural verbs (e) Adjectives (f) Singular proper nouns

(g) Plural proper nouns (h) Function words

Figure 3-12: The average correlation $<Nas>$ of the syntactic representation of the words belonging to the same (in green) or different syntactic categories (in blue). (a) The correlation between the words that belong to singular nouns (*Nsg*) with themselves or with other word categories; likewise, for (b) plural nouns (*Npl*), (c) singular verbs (*Vsg*), (d) plural verbs (*Vpl*), (e) adjectives (*Adj*), (f) singular proper nouns (*Psg*), (g) plural proper nouns (*Ppl*), and (h) function words (*Fwd*). The dashed lines indicate the expected correlations of randomly correlated patterns ($\simeq 3.2$, between content words, and $\simeq 6.4$, between content words and function words).

3.5 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 (Shallice and Cooper, 2011; Shapiro and Caramazza, 2004). We have also made a distinction between the encoding of function words and content words, as suggested by several studies (Friederici and Schoenle, 1980; Friederici et al., 2000). 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.

In order to generate the semantic and syntactic representation of words in a distributed fashion, we have used a generating algorithm that reflects the variable degrees of correlation between words, as suggested by findings on the priming effect (Crescentini et al., 2010; Lerner et al., 2012) and by computational studies for predicting the fMRI signature of words (Mitchell et al., 2008). We observe that a higher semantic dependence (statistical dependence) of words in the BLISS language results in higher correlation in their semantic representation; likewise, a higher syntactical dependence of words results in higher correlation in their syntactic representation. We also observe that the correlations between semantic representation of words within a category is slightly higher than the expected correlations of randomly correlated patterns, and that the correlations between syntactic representation of words within a category is highly above the expected correlations of randomly correlated patterns. To encode a word representation into the Potts network, we have used an auto-associative learning rule—associating a word to itself—for the semantic sub-network; and we have used a hetero-associative—in addition to an auto-associative rule—for the syntactic sub-network, in order to associate different words together through the statistics of word transitions in BLISS.

Our approach toward the representation of words is novel. On the one hand, 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 (Hummel and Holyoak, 1997; Velde and de Kamps, 2006); 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 (Gayler, 2003). 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 (Elman, 1993).

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 (Russo et al., 2010); 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. For the results reported in this chapter, we used a corpus produced by the Subject-Verb semantics model.

Though the Potts network has a large capacity for word storage (Kropff and Treves, 2005), this capacity is valid for randomly correlated—that is, uncorrelated—patterns. For patterns stored using an auto-associative learning rule, computer simulations for the semantic sub-network show that the capacity diminishes for highly correlated patterns (Russo and Treves, 2012). The dependence of the storage capacity on the correlation among stored patterns still needs to be rigorously investigated. Likewise, the storage capacity of the syntactic sub-network—the number of rules that can be stored using a hetero-associative learning rule—needs to be addressed in future studies.

The approach that we implemented here for encoding words of a language of intermediate complexity in a neural network, proposed as a simplified model of the cortex, is a new step toward understanding the principles of word organization in the brain.

Appendix

Table 3.1: Suggestion weight of the syntactic features that served as factors for the generation of function words and different categories of content words.

Table 3.1: Suggestion weights of the syntactic features that served as factors for the generation of function words and different categories of content words. Each element indicates the suggestion weight of a syntactic feature, labelled on the columns, for the generation of a word, labelled on the rows. The column labels are the abbreviation of the syntactic features listed in section 3.3.3. The first 15 row labels list the function words, and the rest are the abbreviation of the lexical categories of the content words.

	lxc/n	lxc/v	lxc/aj	lxc/conj	lxc/prep	lxc/pron	lxc/adv	num/sg	num/pl	neg	det/indf	det/def	det/propn	loc/close	loc/far	dir/from	dir/towards	dir/sameplace	dir/above
thatc	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0.05	0	0.1	0	0
of	0	0	0	0	1	0	0	0	0	0	0	0	0	0.25	0	0.3	0	0	0
in	0	0	0	0	1	0	0.4	0	0	0	0	0	0	0.5	0	0	0	1	0
with	0	0	0	0	1	0	0	0.3	0	0	0	0	0	1	0	0.05	0	0.7	0
on	0	0	0.3	0	1	0	0.4	0	0	0	0	0	0	1	0	0	0	1	1
to	0	0	0	0	1	0	0.1	0	0	0	0	0	0	0	0.5	0	1	0	0
for	0	0	0	0.1	1	0	0	0	0	0	0	0	0	0	0	0	0.1	0	0
doesn't	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0.1	0	0	0	0
don't	0	1	0	0	0	0	0	0.3	1	1	0	0	0	0	0.1	0	0	0	0
the	0	0	1	0	0	0	0	0.5	0.5	0	0	1	0	0	0	0	0	0	0
a	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0
this	0	0	1	0	0	0.5	0	1	0	0	0	1	0	1	0	0	0	0.5	0
that	0	0	1	0	0	0.5	0	1	0	0	0	1	0	0	1	0.25	0.25	0	0
those	0	0	1	0	0	0.5	0	0	1	0	0	1	0	0	1	0.25	0.25	0	0
these	0	0	1	0	0	0.5	0	0	1	0	0	1	0	1	0	0	0	0.5	0
noun/sg	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
noun/pl	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
propn/sg	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0
propn/pl	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0
verb/sg	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
verb/pl	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
adjective	0	0	1	0	0	0	0.3	0	0	0	0	0	0	0	0	0	0	0	0

4

The Potts Network Utters BLISS Words

"Strong Zarathustra leaves Ahriman."

– Potts

Having introduced the Potts network as a simplified model of macroscopic dynamics of the cortex and explained how we encoded the words of the artificial language BLISS in the Potts network, we would like to see how the Potts network behaves with the stored BLISS words.

In this chapter we first illustrate the influence of key parameters of the network by giving some examples of the latching transition using different values for these parameters. Next, we demonstrate the individual behaviour of the semantic and syntactic sub-networks. Finally, we consider the interaction between these two sub-networks, to see what word sequences the Potts network produces.

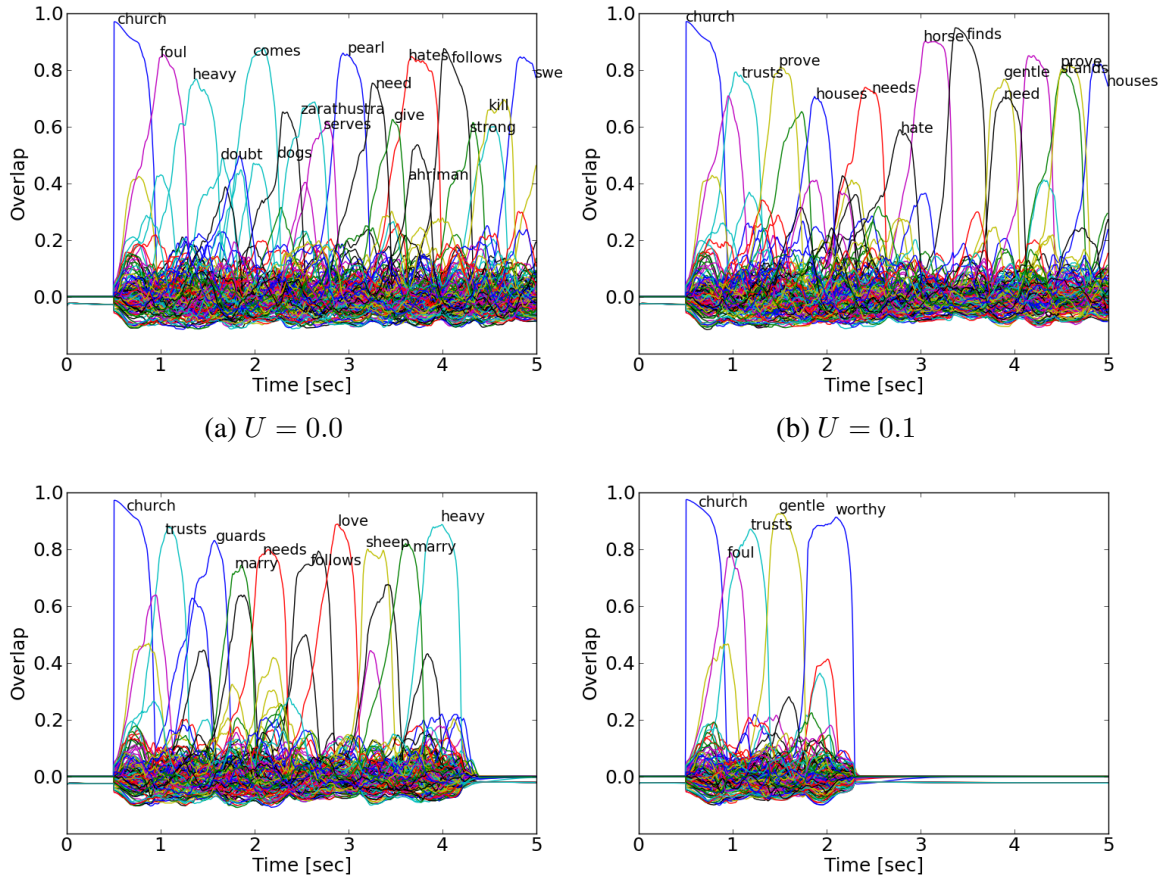
4.1 Effect of the Network Parameters

Before applying the network to the language, we would like to explain the dependence of the latching transitions on the key parameters in the Potts network model. How these parameters influence the latching transition in the network was extensively investigated by (Russo and Treves, 2012), who showed the phases in which the finite and infinite latching exist. Here, we visualize this influence through concrete examples.

We stored $p = 149$ patterns using the auto-associative learning rule (3.2) on $N = 541$, using *randomly* generated patterns ($< Nas > \simeq 4.2$). The number of connections per unit was set around $C = 88$, keeping the connections diluted at $C/N \approx 0.16$. With this network, we investigated the role of U (the fixed threshold), w (the local feedback term), β (the inverse temperate), and τ_1, τ_2, and τ_3 (the time constants of the dynamic thresholds).

4.1.1 The Fixed Threshold: U

As shown in Eq. 3.11, the constant U is the input to the null state, in other words, it is the constant threshold that an active state must surpass to exceed the activation level of the null state. The higher U, the higher the activation level of the null state, the more difficult for an active state to be the winner ($\sum_{k=0}^{S} \sigma_i^k(t) = 1$), thus the more probable for the network to stop latching. This influence is visualized in Fig. 4-1.



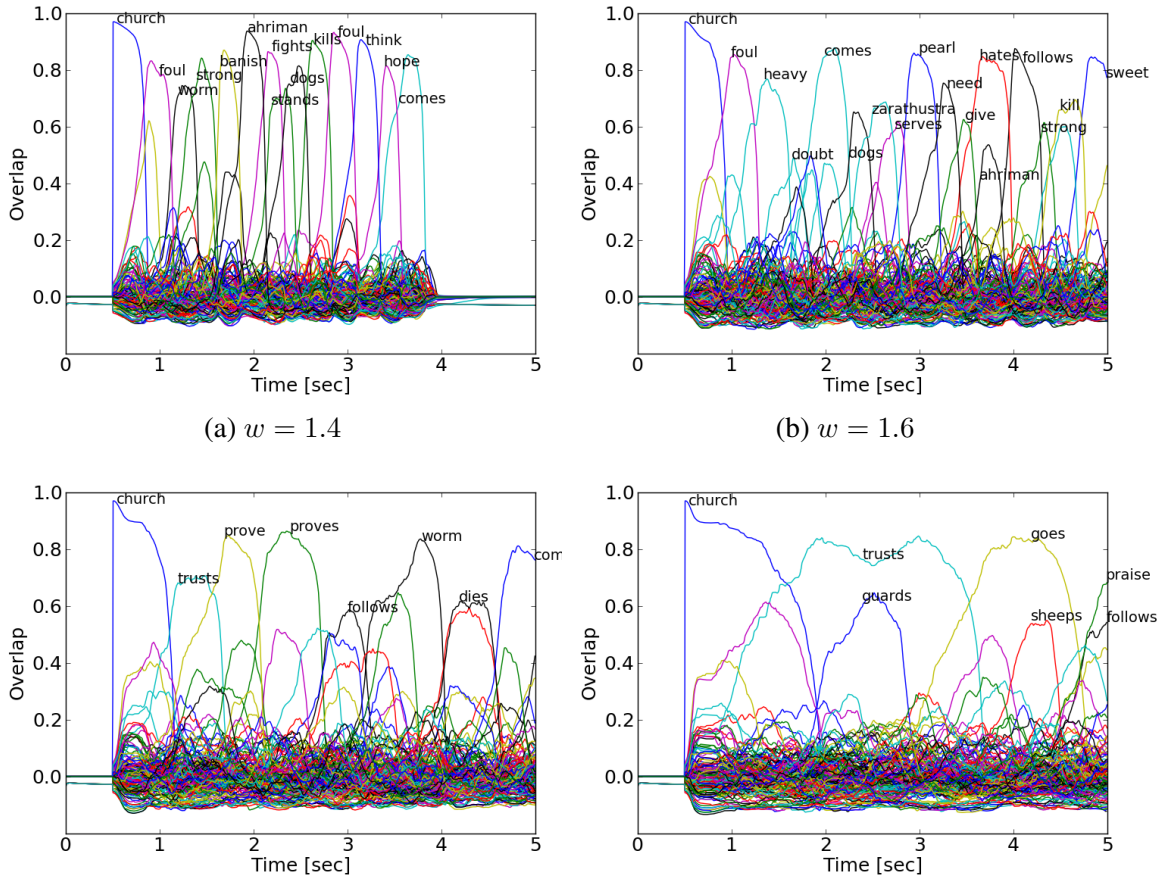
(a) $U = 0.0$ (b) $U = 0.1$

(c) $U = 0.2$ (d) $U = 0.3$

Figure 4-1: Effect of varying the fixed threshold U in the Potts network with randomly correlated patterns: (a) $U = 0.0$, (b) $U = 0.1$, (c) $U = 0.2$, and (d) $U = 0.3$. The x axis is time and the y axis is the correlation (overlap) of the network state with each of the stored patterns, or words, indicated in different colours. The parameters were set at $N = 541$, $p = 149$, $C = 88$, $\tau_1 = 10$, $\tau_2 = 200$, $\tau_3 = 10000$, $w = 1.6$, and $\beta = 5$.

4.1.2 The Self-reinforcement Term: w

The feedback term or self-reinforcement term, w (Eq. 3.13), provides a positive feedback to an active state of a unit if the activation level of the state is relatively higher than the average activity across active states of the unit; therefore, a higher w keeps active states of high activity more stable. As shown in Fig. 4-2, the number of transitions decreases as w increases, since a retrieved pattern tends to remain active for a longer time (due to strong reinforcement of its active states).



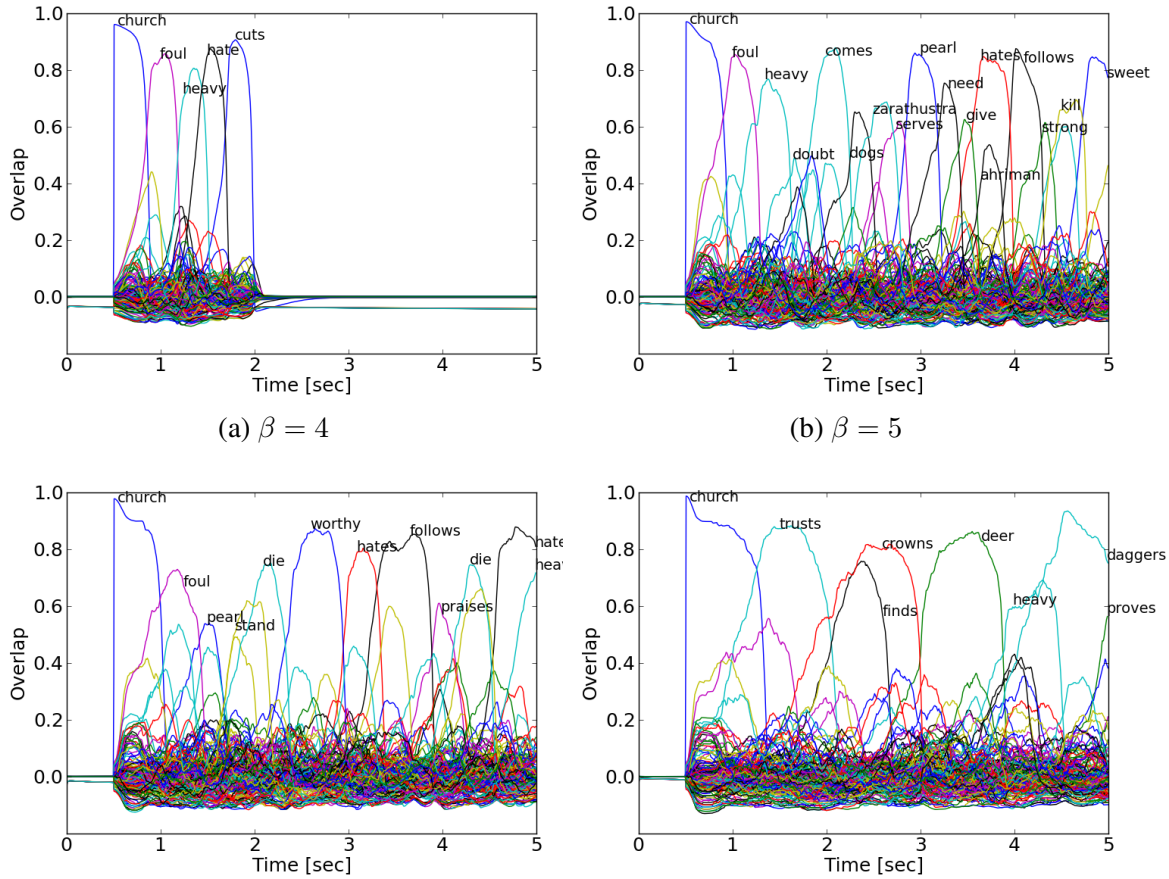
(a) $w = 1.4$ (b) $w = 1.6$

(c) $w = 1.8$ (d) $w = 2.0$

Figure 4-2: Effect of varying the self-reinforcement term, w, in the Potts network with randomly correlated patterns: (a) $w = 1.4$, (b) $w = 1.6$, (c) $w = 1.8$, and (d) $w = 2.0$. The x axis is time and the y axis is the correlation (overlap) of the network state with each of the stored patterns, or words, indicated in different colours. The parameters were set at $N = 541$, $p = 149$, $C = 88$, $\tau_1 = 10$, $\tau_2 = 200$, $\tau_3 = 10000$, $U = 0.0$, and $\beta = 5$.

4.1.3 The Inverse Temperature: β

The inverse of the thermal noise, β, in the Potts model affects the activation function of a state as a factor in the exponent of the exponential function (Eqs. 3.10 and 3.11), thus adjusting the sensitivity of the activation function of a state to changes of its firing rate. High noise (small β) attenuates activity changes and does not allow highly active states to have a voice; consequently, the network more probably dies out, as shown in Fig. 4-3. In contrast, low noise (high β) bestows more stability upon the retrieved patterns, by increasing the sensitivity of the activation function to firing rate changes.



(a) $\beta = 4$ (b) $\beta = 5$

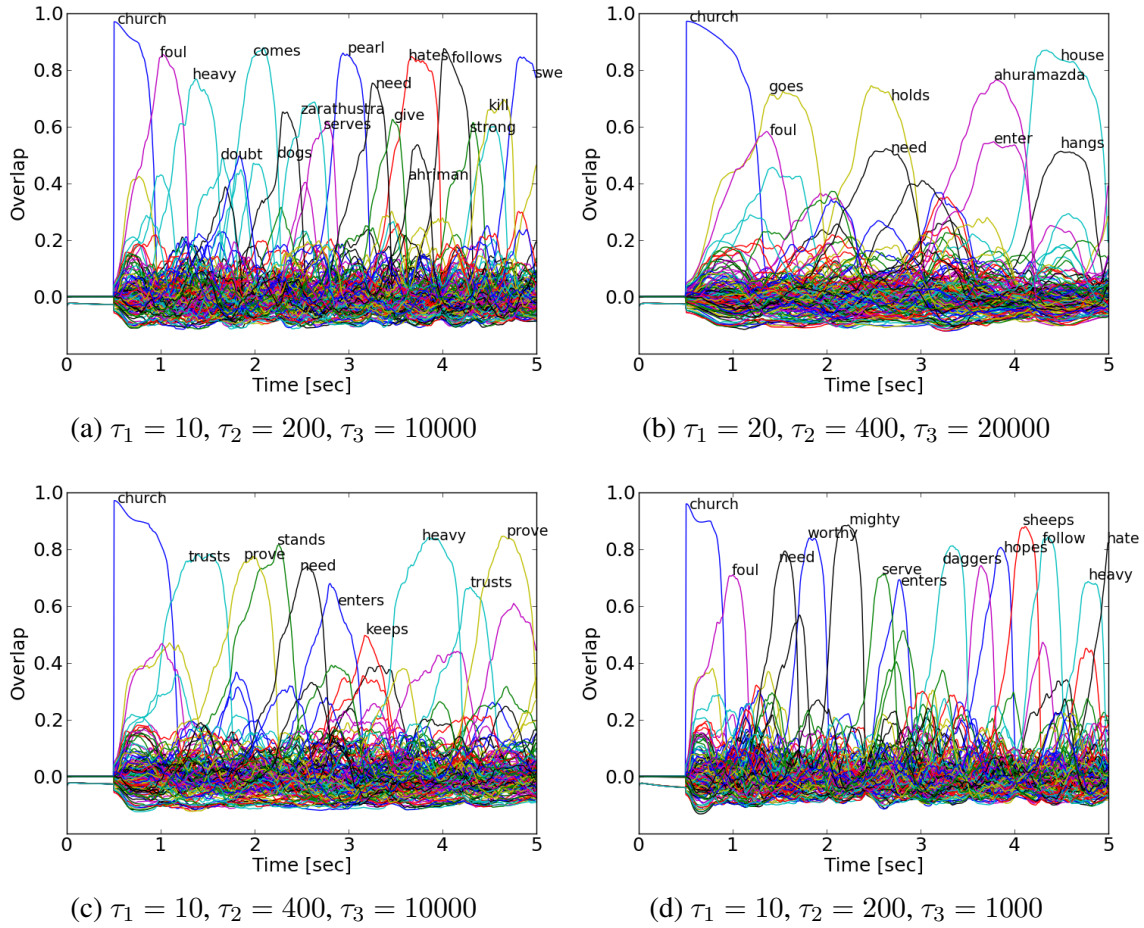
(c) $\beta = 6$ (d) $\beta = 8$

Figure 4-3: Effect of varying the inverse temperature, β, in the Potts network with randomly correlated patterns: (a) $\beta = 4$, (b) $\beta = 5$, (c) $\beta = 6$, and (d) $\beta = 8$. The parameters were set at $N = 541$, $p = 149$, $C = 88$, $\tau_1 = 10$, $\tau_2 = 200$, $\tau_3 = 10000$, $U = 0.0$, and $w = 1.6$.

4.1.4 The Time Constants: τ_1, τ_2, τ_3

As discussed in the previous chapter, dynamic thresholds are introduced to the dynamics of Potts units to model neuronal fatigue and slow inhibition in cortical patches. The time constants of these threshold dynamics—τ_1, τ_2, and τ_3—determine, respectively, the speed of firing rate integration, $r_i^k(t)$, the speed of variation of threshold of active states, $\theta_i^k(t)$, and the speed of variation of threshold of the null state, $\theta_i^0(t)$. Compare Fig. 4-4a and Fig. 4-4b to see that we can slow down the latching transitions by increasing these time constants by a common factor. Keeping the time constants in Fig. 4-4a as a reference, we note in Fig. 4-4c that the network stays longer in an attractor even when increasing only τ_2 (which slows down the decay of the activation of an active state). In Fig. 4-4d,

we see that the network shows more clear transitions between the dominant patterns while inhibiting the competitive patterns, when decreasing τ_3 (which speeds up the dynamics of the common threshold).



(a) $\tau_1 = 10$, $\tau_2 = 200$, $\tau_3 = 10000$

(b) $\tau_1 = 20$, $\tau_2 = 400$, $\tau_3 = 20000$

(c) $\tau_1 = 10$, $\tau_2 = 400$, $\tau_3 = 10000$

(d) $\tau_1 = 10$, $\tau_2 = 200$, $\tau_3 = 1000$

Figure 4-4: Effect of varying the time constants of the dynamic thresholds, τ_1, τ_2, and τ_3, in the Potts network with randomly correlated patterns. Keeping the time constants of (a) as a reference, in (b) the time constants are doubled, in (c) only τ_2 is increased, and in (d) τ_3 is decreased. The parameters were set at $N = 541$, $p = 149$, $C = 88$, $\beta = 5$, $U = 0.0$, and $w = 1.6$.

As discussed in this section, different ranges of network parameters result in different speeds for the latching transitions in the network. Given an average human speaking rate of about 150 words per minute, we decided to choose the following values of the parameters for subsequent simulations: $\tau_1 = 10$, $\tau_2 = 200$, $\tau_3 = 10000$, $U = 0.0$, $\beta = 5$, and $w = 1.6$, which result in about 11 words in 4.5 seconds.

4.2 The Potts Semantic Network

In the preceding chapter we explained how we generated the semantic representation of 149 words (section 3.3.2), and we described how to encode the words into the semantic sub-network using an auto-associative learning rule. Here we want to examine the dynamics of the semantic sub-network with these stored words (section 3.3.4).

Fig. 4-5 shows the behaviour of the semantic sub-network after encoding words into it. We observe the dominance of the words *house* and *give* (plural and singular forms were represented identically) in the network dynamics, a consequence of the underlying correlation between these words. Though the average correlation among words in the semantic sub-network is close to that of randomly correlated pattern, some pairs of words exhibit correlations that highly exceed the average.

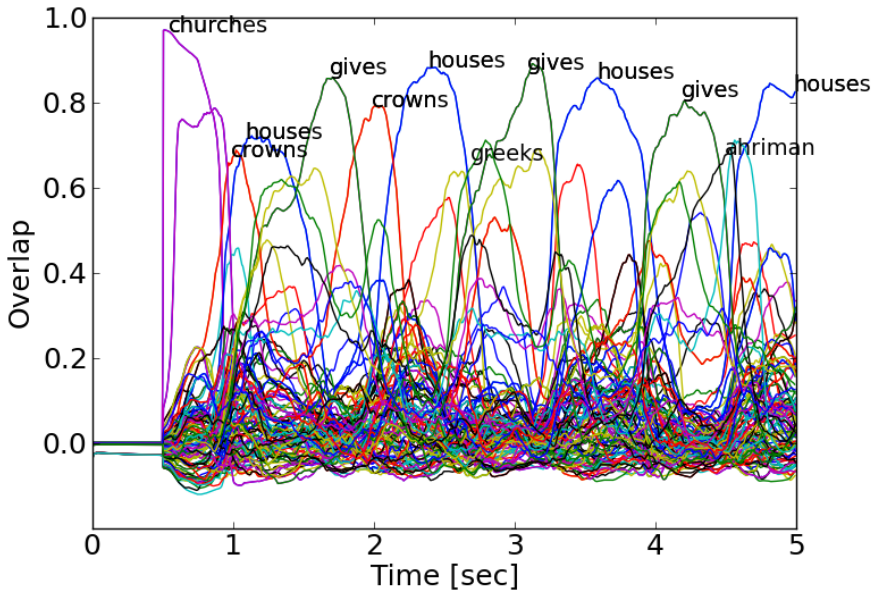


Figure 4-5: **Limit cycle dynamics with correlated patterns:** latching dynamics in the semantic sub-network with words encoded by the auto-associative learning rule and generated as explained in the previous chapter. The parameters were set at $N_{sem} = 541$, $C_{sem} = 88$, $\tau_{1\,sem} = 10$, $\tau_{2\,sem} = 200$, $\tau_{3\,sem} = 10000$, $w_{sem} = 1.6$, $\beta_{sem} = 5$, and $U_{sem} = 0.0$.

Unfortunately given the procedure chosen here for encoding words in the semantic sub-network, we were not able to find a simple heuristic approach for escaping this limit cycle phase. The phase diagram of the latching network in the presence of some larg correlation

needs to be investigated by future studies.

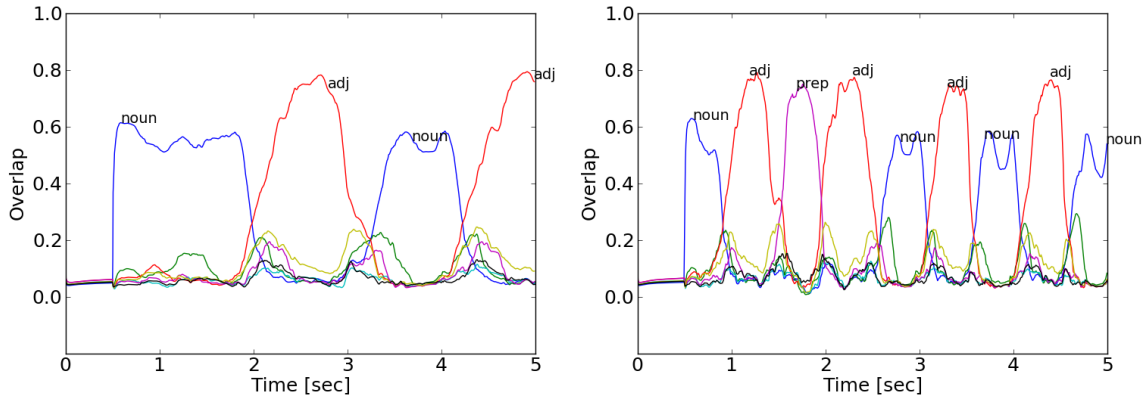
4.3 The Potts Syntactic Network

In the preceding chapter, we explained how to generate the syntactic representation of 149 words (section 3.3.3) and we described the way to encode the words into the syntactic sub-network using both auto-associative and hetero-associative learning rules (section 3.3.4). Here we want to examine the dynamics of the syntactic sub-network with these stored words, for $N = 359$ and $C = 58$ ($C/N \approx 0.16$).

To visualize the correlation (overlap) between the network activity and the 149 stored words, we decided to show the correlation not with all the words but with the 7 lexical categories (section 3.3.3) used as the factors for generating the syntactic representation of words: lxc/noun, lxc/verb, lxc/adj, lxc/conjunction, lxc/preposition, lxc/pronoun (for simplicity "lxc/" is dropped from their names). We made this choice because in the syntactic sub-network, the overlap of the words within a category is very high (Fig. 3-12); instead, we use these 7 lexical categories as indicators of different word categories (Fig. 3-11).

4.3.1 Latching with the Auto-associative Learning Rule Alone

In Fig. 4-6 we observe the behaviour of the syntactic sub-network with the words stored by only using the auto-associative learning rule ($c_{auto} = 1.0$ and $c_{hetero} = 0.0$ in Eq. 3.1). The transitions in the syntactic sub-network are slow (Fig. 4-6a) when using the time constants of the dynamic thresholds found appropriate for the case of randomly correlated patterns (section 4.1); to compensate, we doubled the speed by halving the time constants (Fig. 4-6b). This difference in the speed might originate from the small number of patterns that we effectively stored in the syntactic sub-network compared to the semantic sub-network; due to the high correlations between words in the syntactic sub-network, the patterns that were effectively stored is about the size of the lexical categories.

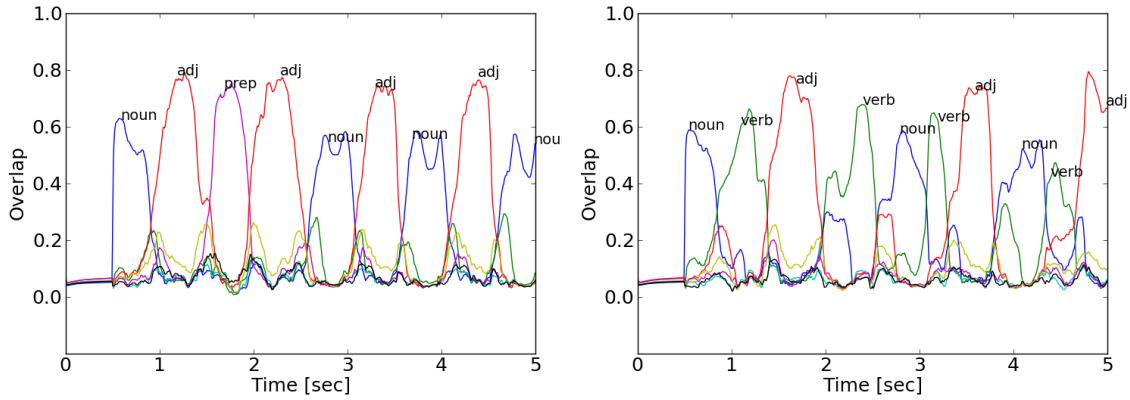


(a) $\tau_{1\,syn} = 10$, $\tau_{2\,syn} = 200$, $\tau_{3\,syn} = 10000$ (b) $\tau_{1\,syn} = 5$, $\tau_{2\,syn} = 100$, $\tau_{3\,syn} = 5000$

Figure 4-6: Dynamics of the syntactic sub-network with (a) slower and (b) faster time constants. The patterns were encoded by an auto-associative learning rule ($c_{auto} = 1.0$ and $c_{hetero} = 0.0$, in Eq. 3.1) and were generated by the procedure discussed in the previous chapter. The x axis is time and the y axis is the correlation (overlap) of the network state with 7 word categories used as the factors for the syntactic representation of the words. The parameters were set at $N_{sem} = 359$, $C_{syn} = 58$, $U_{syn} = 0.0$, $w_{syn} = 1.6$, $\beta_{syn} = 5$.

4.3.2 Turning on the Hetero-associative Learning Rule

Having adjusted the time constants in the syntactic sub-network, we now turn on the hetero-associative learning rule ($c_{hetero} \neq 0.0$) while keeping $c_{auto} = 1.0$. As we see in Fig. 4-7, introducing the hetero-associative component gives the network some knowledge about the possible sequences of words and leads the transitions toward more grammatically valid sequences. For $c_{hetero} = 0.1$ (Fig. 4-7b), the valid sequences form "noun verb" (0.5–1.5 sec), "noun verb adj noun" (3.0–4.0 sec), and "adj noun verb" (3.5–4.5 sec). However, we observe in invalid sequences like "adj verb" (1.5–2.5 sec), where *adj* retrieves both *noun* and *verb* (2.0 sec), yet *noun* loses this competition, while it successfully wins it later (4.0 sec). The outcome of this competition does not seem to be improved by increasing c_{hetero} (Figs. 4-7c and 4-7d).



(a) $c_{hetero} = 0.0$ (b) $c_{hetero} = 0.1$

(c) $c_{hetero} = 0.2$ (d) $c_{hetero} = 0.3$

Figure 4-7: **Hetero-associative learning introduces syntax:** latching dynamics in the syntactic sub-network with the words encoded using both an auto-associative learning rule with fixed $c_{auto} = 1.0$, and a hetero-associative learing rule with (a) $c_{hetero} = 0.0$, (b) $c_{hetero} = 0.1$ (c) $c_{hetero} = 0.2$, and (d) $c_{hetero} = 0.3$. The parameters were set as $C_{syn} = 58$, $\tau_{1\,syn} = 5$, $\tau_{2\,syn} = 100$, $\tau_{3\,syn} = 5000$, $w_{syn} = 1.6$, $\beta_{syn} = 5$, $U_{syn} = 0.0$.

4.4 Interaction of Semantic and Syntactic sub-networks

We coupled the semantic and syntactic sub-networks to each other to observe whether they cooperate to produce interesting sentences. These two sub-networks influence each other by different weights (Eq. 3.19): $c_{sem2syn}$, the weight by which the semantic sub-network influences the syntactic sub-network, and $c_{syn2sem}$, the weight by which the syntactic sub-network influences the semantic sub-network. This interaction influences the number of connections per unit C, as in Eq. 3.21.

4.4.1 The Semantic sub-network with Correlated Patterns

As we observed earlier, in section 4.2, if we try to use the semantic sub-network with words generated and encoded through the procedure explained in the last chapter, the sequences were uninterestingly composed of the two words *house* and *give*. If we allow this sub-network to be influenced by the syntactic sub-network (Fig. 4-8), we still see the dominance of these two words. The sentences, both for semantic and syntactic retrievals, are shown at the top of the figure and roughly matched by their co-activations. Note several incorrect retrievals marked by *; these were not guided by the syntactic sub-network. Despite this dominance and the errors, we see that the syntactic sub-network attempts to retrieve relevant words, though with smaller overlap compared to *house* and *give*; for instance, *adj* in the syntactic network (1.7 sec) brings up *poor* in the semantic part; similarly, *noun* (3.0 sec) evokes *Greeks* and later (4.0 sec) it retrieves *crowns*.

Because of this unsatisfactory behaviour, we decided to work with randomly correlated patterns for the semantic network, as in section 4.1. In what follows, we focus on a semantic sub-network with randomly correlated patterns.

churches	*houses	*gives	crowns	*houses	houses	gives	*houses	crowns	gives	*houses
noun	verb	adj	[noun]	verb	noun	verb	adj	noun	verb	adj

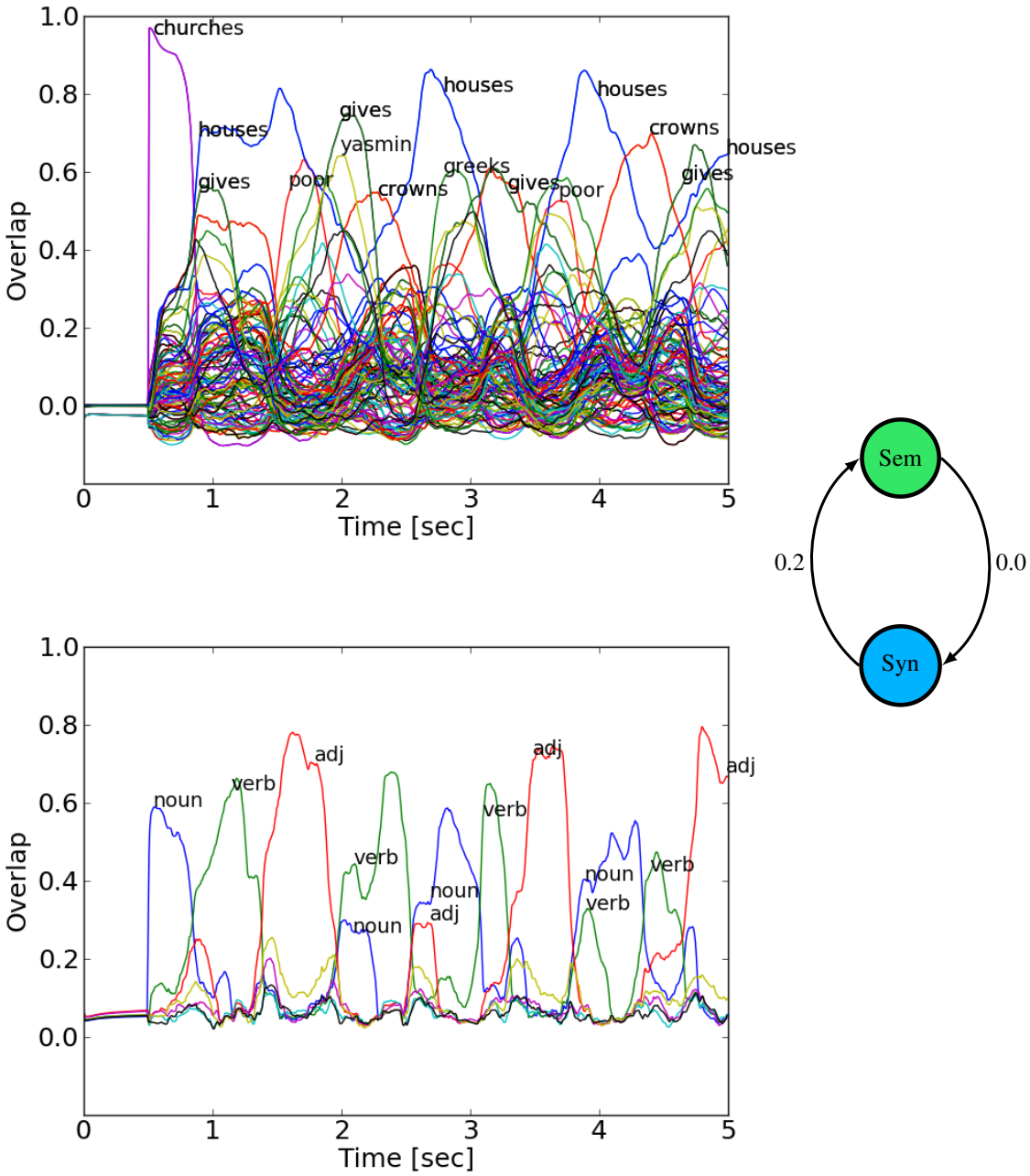


Figure 4-8: **Correlated patterns are difficult to harness:** the interaction between the semantic sub-network with correlated patterns (top) and the syntactic sub-network (bottom). The strength of the influence of the syntactic part on the semantic part is $c_{syn2sem} = 0.2$ and of the semantic part on the syntactic part is $c_{sem2syn} = 0.0$, see right panel. The sentences produced by the interaction of these two sub-networks are shown at the top; incorrect retrievals are marked by * and incomplete valid retrievals are included in []. The patterns in the syntactic sub-networks are generated with both auto- and hetero-associative learning, $c_{auto} = 1.0$ and $c_{hetero} = 0.1$. The parameters were set at $w = 1.6$, $\beta = 5$, $U = 0.0$; for the semantic sub-network, $C_{sem} = 100$, $\tau_{1\,sem} = 10$, $\tau_{2\,sem} = 200$, $\tau_{3\,sem} = 10000$; for the syntactic sub-network, $C_{syn} = 75$, $\tau_{1\,syn} = 5$, $\tau_{2\,syn} = 100$, $\tau_{3\,syn} = 5000$.

4.4.2 The Semantic sub-network Influences the Syntactic sub-network

We now allow the semantic sub-network to guide the syntactic one through $c_{sem2syn} = 0.1$. As shown at the top of Fig. 4-9, some retrievals in the syntactic sub-network correctly follow the corresponding active word in the semantic part; for example, *foul* and *heavy* in the semantic part (1.0–1.5 sec) correctly induce the retrieval of *adj* in the syntactic part; likewise, *doubt* and *comes* (1.5–2.0 sec) induce the retrieval of *verb*. On the other hand, there are several instances of invalid retrievals, marked by * in the sentences, or incomplete retrievals, marked by []: for example, *comes* (2.0 sec) incorrectly retrieves *noun*, and *prove* (3.0 sec) incorrectly retrieves *noun*, while the word *foolish* (4.5 sec) does not succeed to completely induce the retrieval of *adj*.

Increasing the influence of the semantic sub-network on the syntactic sub-network enhances by some, but not enough, the correspondence between semantic and syntactic patterns (Fig. 4-10). The sentences written on the top shown that most of the retrieved words in the syntactic pattern correctly match the corresponding retrieved words in the semantic part; however, *hates* incorrectly provoked *noun* (3.8 sec) and *strong* also incorrectly provoked *noun* (4.2 sec).

4.4.3 The Syntactic sub-network Influences the Semantic sub-network

We would now like to examine the influence of the syntactic on the semantic sub-network. Fig. 4-11 shows the guidance of the semantic part by the syntactic sub-network through $c_{syn2sem} = 0.1$. The match between the two sub-networks is not perfect. Although in most cases the retrieved words in the semantic part match their word category in the syntactic network, there are still a few cases of incongruence—*know* was incorrectly retrieved by *adj* (3.7 sec), and *Zarathustra* by *verb* (4.4 sec).

If we increase the influence of the syntactic sub-network, we see much more interesting and grammatically valid sentences (Fig. 4-12). Though we have not yet introduced punctuations (e.g. period or exclamation mark for distinguishing the sentences) or coordinating conjunctions (e.g. *and*, *or*), the network roughly produced interesting sentences such as:

- Church trusts [and] holds sweet, heavy Zarathustra!

- Sweet, heavy Zarathustra holds pearl.
- Pearl stands.
- Heavy Zarathustra [or] pearl stands.

(*Zarathustra* was the founder and prophet of Zoroastrianism—an ancient Persian religion.)

The semantic network tends to produce more than one word for a word category proposed by the syntactic network, such as ”trusts holds” for *verb* or ”sweat heavy” for *adj*. So as to help the semantic network listen more clearly to the syntactic part, we decreased the talking speed of the semantic network by increasing the time constants of the dynamic thresholds. Fig. 4-13 shows how effective the decrease of the time constants are; the semantic sub-network speaks one word of the corresponding word category, instead of quickly uttering two words:

- Church holds strong Zarathustra!
- Strong Zarathustra leaves Ahriman!
- Ahriman trusts heavy gate.
- gate takes foolish . . .

(*Ahriman* indicates a sort of devil in Zoroastrianism.)

church noun	foul, heavy adj	doubt, comes verb	comes *noun	dies, stands verb	prove *noun	prove verb	crowns noun	doubts verb	hate *noun	hate verb	foolish [adj]	foolish *noun	finds verb

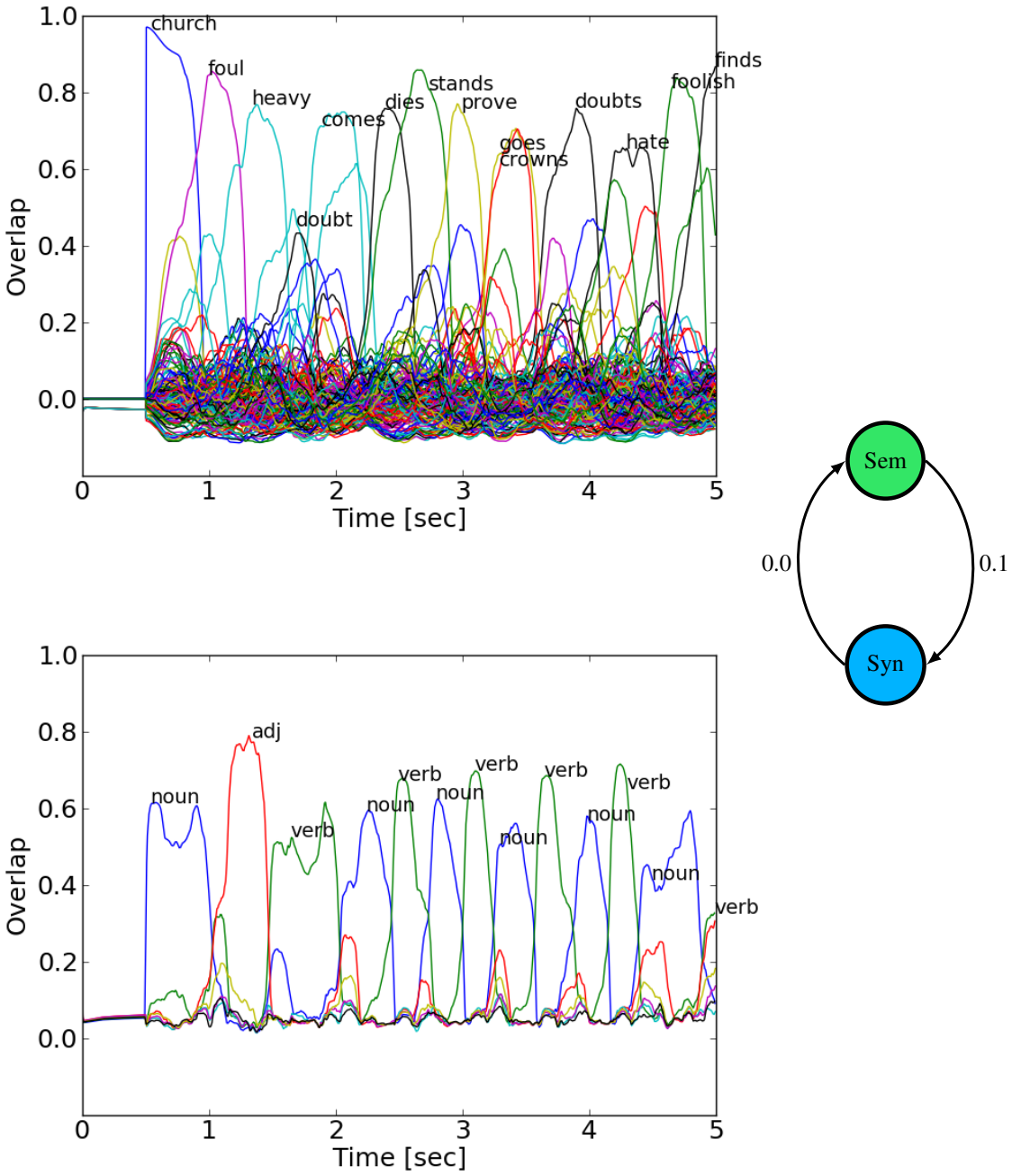


Figure 4-9: **Semantics is a loose guide for syntax, if the connections are weak:** the semantic sub-network with randomly correlated patterns (top) influences the syntactic sub-network (bottom) with the weight $c_{sem2syn} = 0.1$. The sentences produced by the interaction of these two sub-networks are written on the top; incorrect retrievals are marked by * and incomplete valid retrievals are included in []. The parameters were set at $w = 1.6$, $\beta = 5$, $U = 0.0$; for the semantic sub-network, $C_{sem} = 88$, $\tau_{1\,sem} = 10$, $\tau_{2\,sem} = 200$, $\tau_{3\,sem} = 10000$; for the syntactic sub-network, $C_{syn} = 68$, $\tau_{1\,syn} = 5$, $\tau_{2\,syn} = 100$, $\tau_{3\,syn} = 5000$, $c_{auto} = 1.0$, and $c_{hetero} = 0.0$

church noun	foul, heavy adj	doubt, comes verb	dogs, Zarathustra noun	serves verb	pearl noun	need, give verb	hates *noun	follows verb	strong *noun	kill verb	sweet *noun

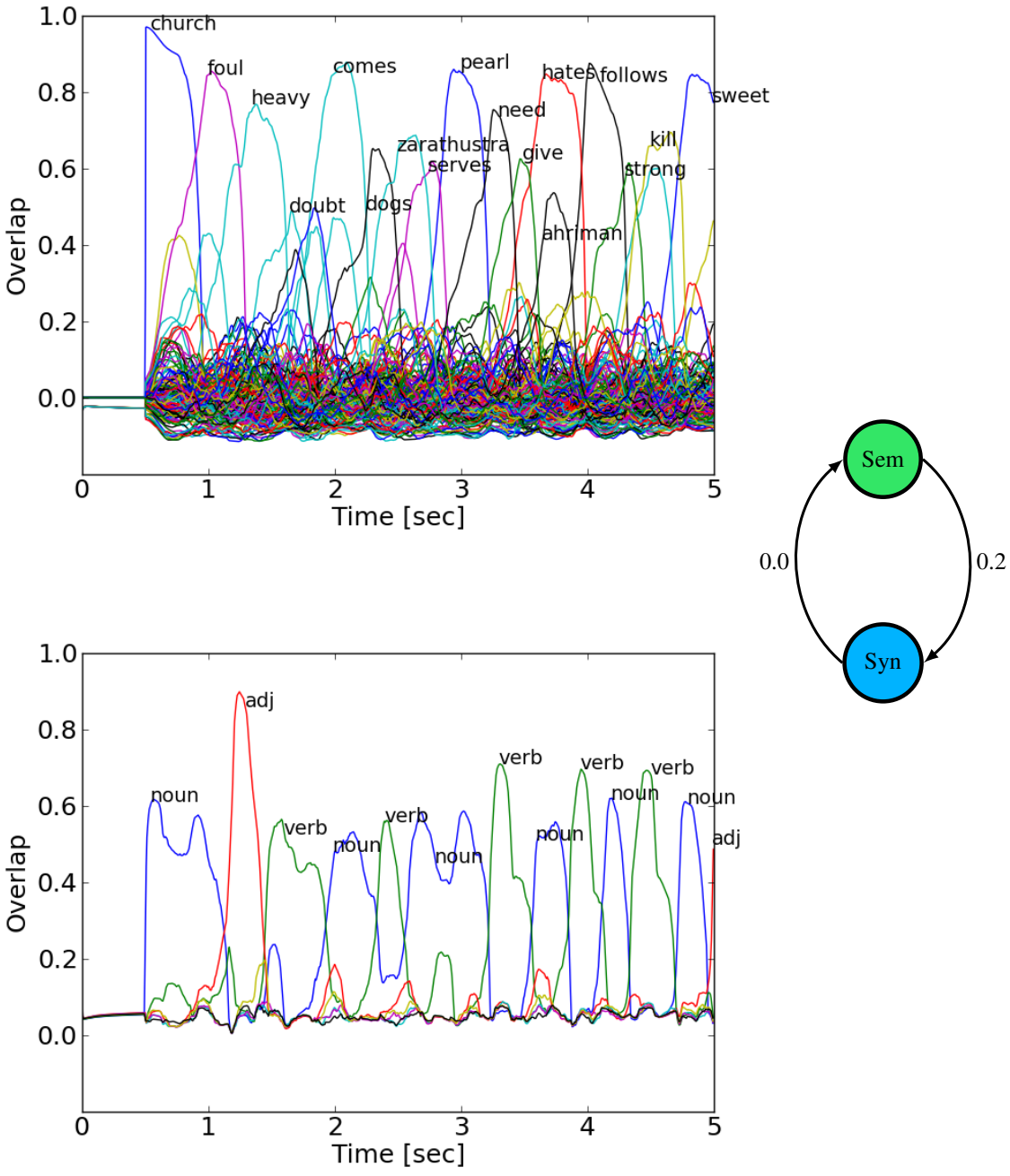


Figure 4-10: **A stronger coupling does not improve much the correspondence:** the semantic sub-network (with randomly correlated patterns) (top) influences the syntactic sub-network (bottom) with the weight $c_{sem2syn} = 0.2$. The sentences produced by the interaction of these two sub-networks are written on the top. The parameters were set at $w = 1.6$, $\beta = 5$, $U = 0.0$; for the semantic sub-network, $C_{sem} = 88$, $\tau_{1\,sem} = 10$, $\tau_{2\,sem} = 200$, $\tau_{3\,sem} = 10000$; for the syntactic sub-network, $C_{syn} = 75$, $\tau_{1\,syn} = 5$, $\tau_{2\,syn} = 100$, $\tau_{3\,syn} = 5000$, $c_{auto} = 1.0$, and $c_{hetero} = 0.0$.

church, Zarathustra	marries, enters	foul	dies, need	Ahriman, horse	wear	*know	stands	sword	*Zarathustra	worthy
noun	verb	adj	verb	noun	verb	adj	[verb]	noun	verb	adj

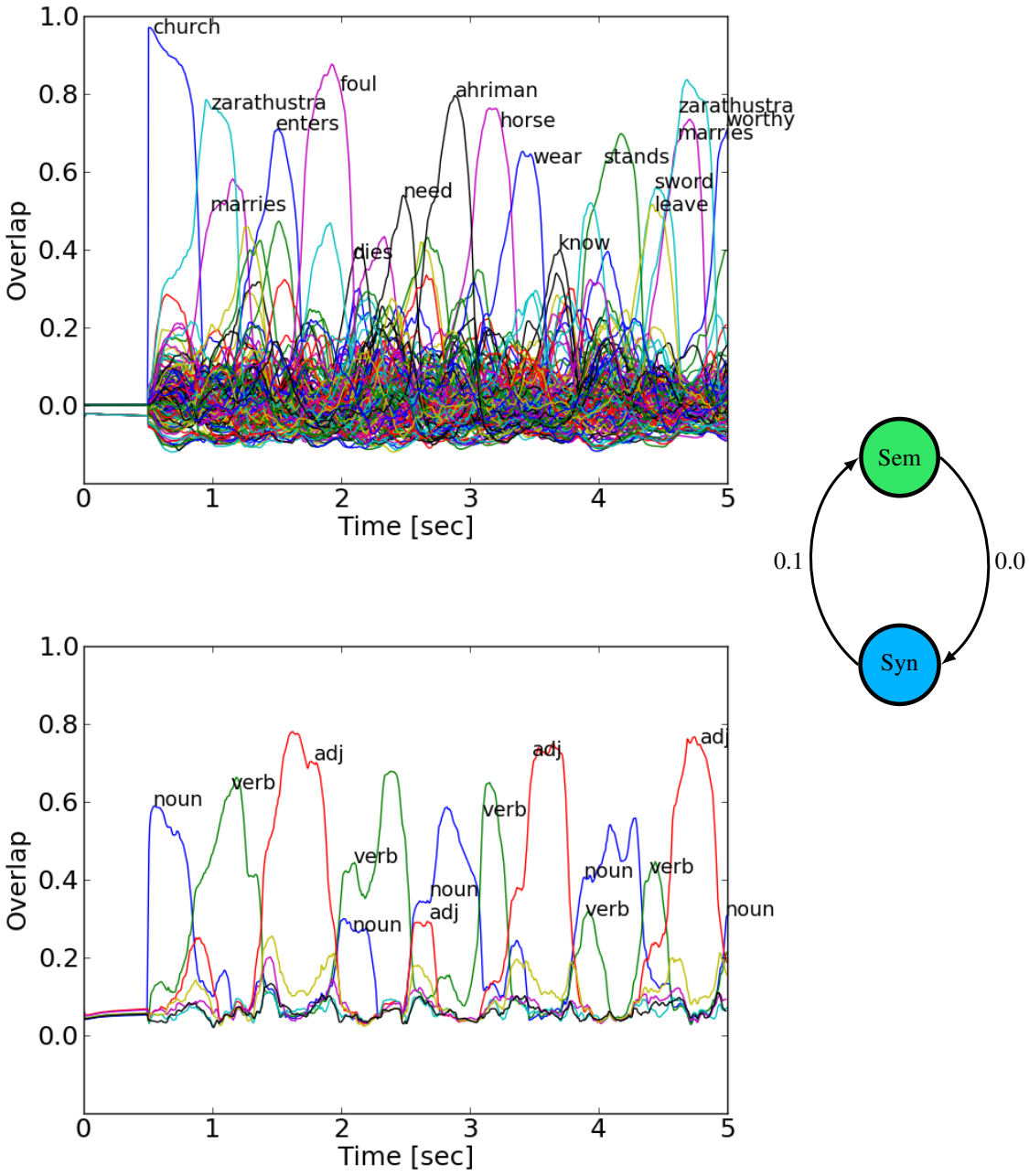


Figure 4-11: **A weak coupling from syntax to semantics is still ineffective:** the syntactic sub-network (bottom) influences the semantic sub-network (top) with the weight $c_{syn2sem} = 0.1$. The sentences produced by the interaction of these two sub-networks are written on the top. The parameters were set at $w = 1.6$, $\beta = 5$, $U = 0.0$; for the semantic sub-network, $C_{sem} = 94$, $\tau_{1\,sem} = 10$, $\tau_{2\,sem} = 200$, $\tau_{3\,sem} = 10000$; 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 $c_{hetero} = 0.1$.

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

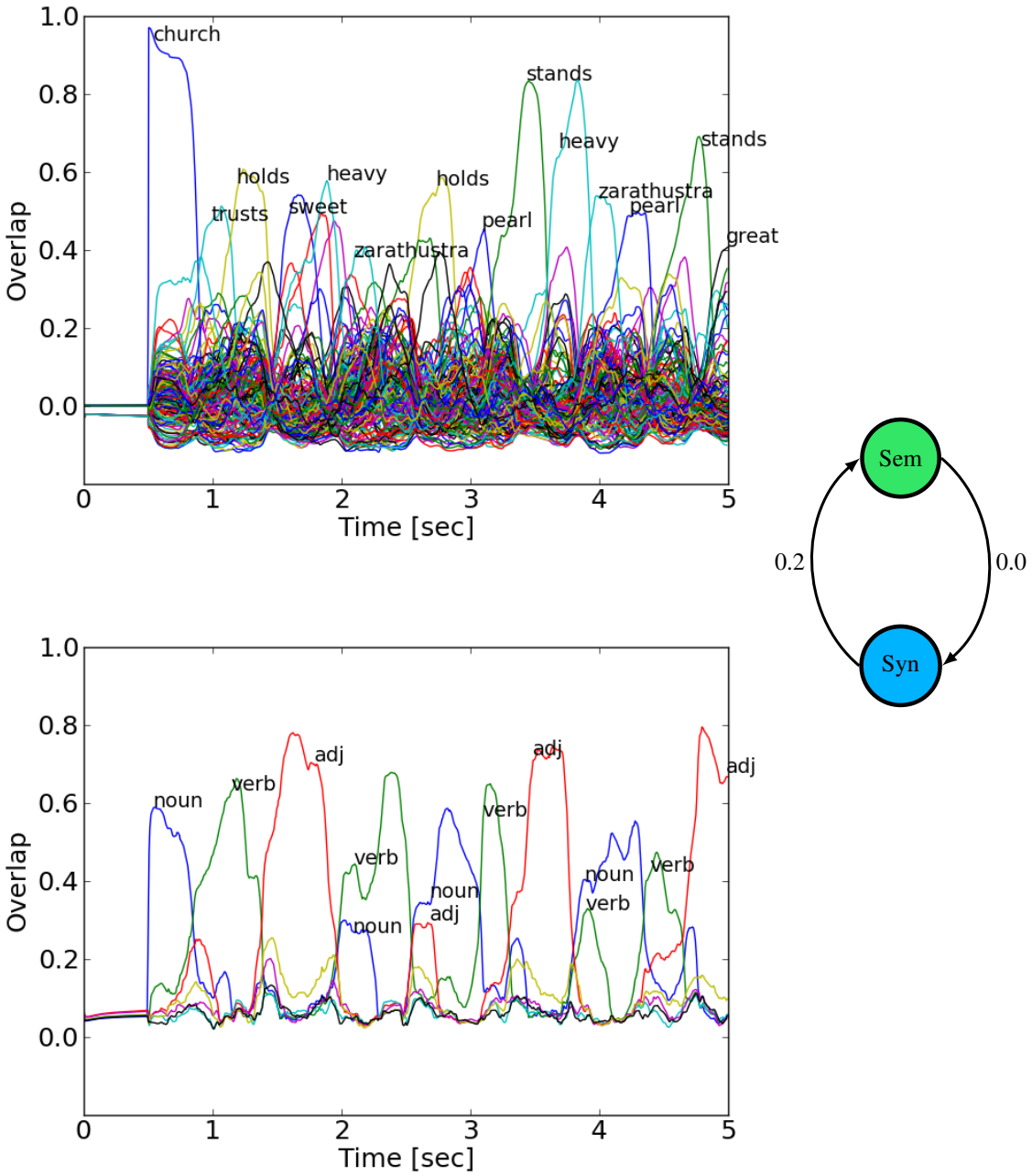


Figure 4-12: **A stronger coupling now improves the match:** the syntactic sub-network (bottom) influences the semantic sub-network (top) with the weight $c_{syn2sem} = 0.2$. The sentences produced by the interaction of these two sub-networks are written on the top. The parameters were set at $w = 1.6$, $\beta = 5$, $U = 0.0$; for the semantic sub-network, $C_{sem} = 100$, $\tau_{1\,sem} = 10$, $\tau_{2\,sem} = 200$, $\tau_{3\,sem} = 10000$; 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 $c_{hetero} = 0.1$.

church	holds	strong	Zarathustra	leaves	Ahriman	trusts	heavy	gate	takes	foolish
noun	verb	adj	[noun]	verb	noun	verb	adj	noun	verb	adj

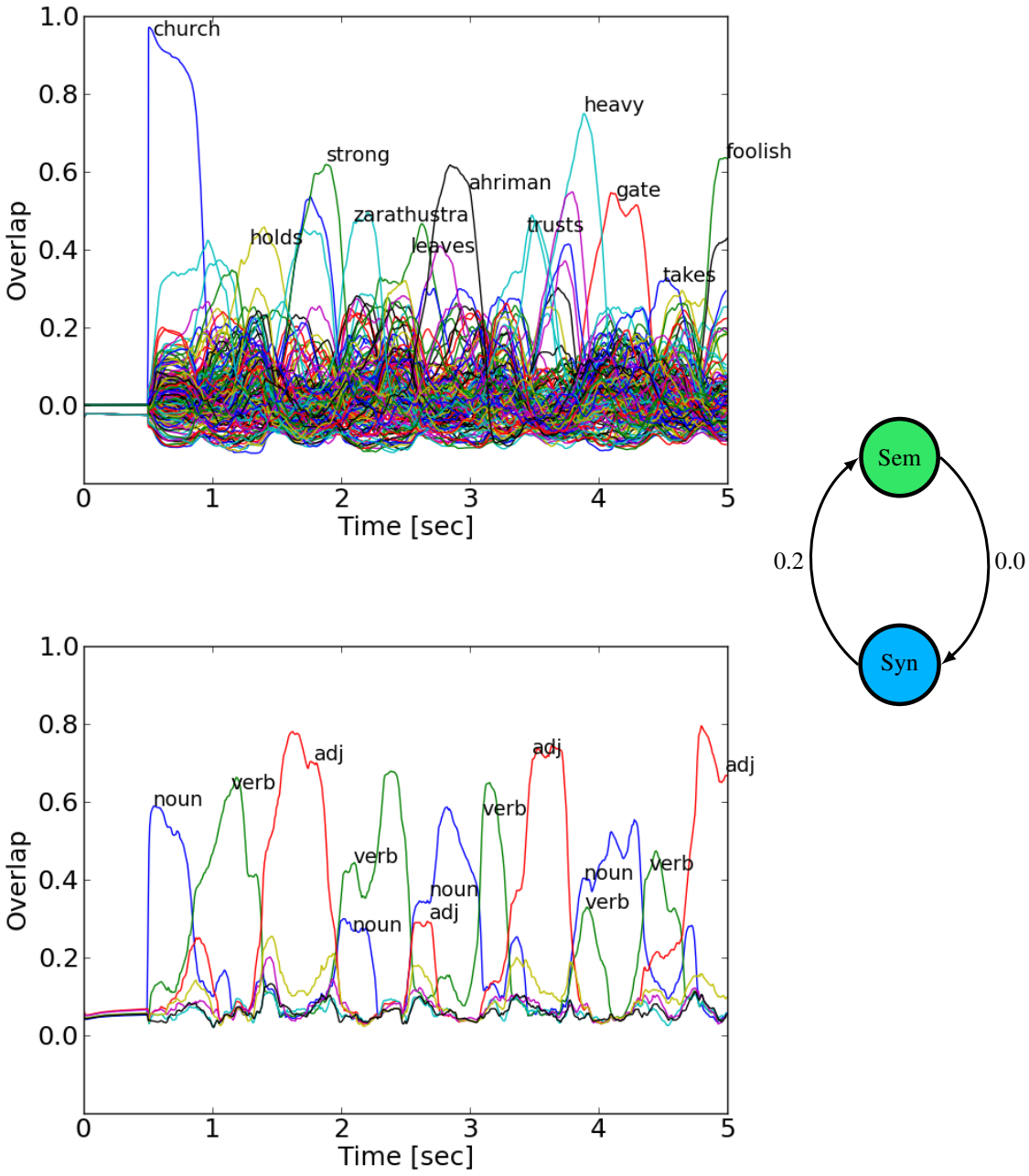


Figure 4-13: **Further improvement with slower semantics dynamics:** the syntactic sub-network (bottom) influences the semantic sub-network (top) with the weight $c_{syn2sem} = 0.2$, while the semantic sub-network moves slowly. The parameters were set at $w = 1.6$, $\beta = 5$, $U = 0.0$; for the semantic sub-network, $C_{sem} = 100$, $\tau_{1\,sem} = 12.5$, $\tau_{2\,sem} = 250$, $\tau_{3\,sem} = 12500$; 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 $c_{hetero} = 0.1$.

4.4.4 The Semantic sub-network and the Syntactic sub-network Co-operate

Having investigated the directed interaction between the sub-networks, semantic to syntactic and syntactic to semantic), we now examine their bi-directional cooperation, with $c_{sem2syn} = c_{syn2sem} = 0.1$, in Fig. 4-14. We compare this bi-directional interaction with the uni-directional influence of the semantic on the syntactic network (Fig. 4-9), where mostly the verbs were active in the semantic part and these verbs produced a consecutive repetition of nouns and verbs in the syntactic part. Similarly (Fig. 4-14), we now see that the syntactic part influences the semantic part to produce several repetitions of nouns and verbs (i.e "church, Zarathustra; holds, doubts; church, Zarathustra; holds, dare; pearl, sword;").

For a stronger bidirectional interaction (Fig. 4-15), we see an even stronger consecutive repetition of nouns and verbs, although we also see better cooperation, with much less discrepancy between the two networks.

church, Zarathustra noun	holds, doubts verb	church, Zarathustra noun	holds, dare verb	pearl, sword noun	*Zarathustra adj	worthy adj	hate, leaves verb

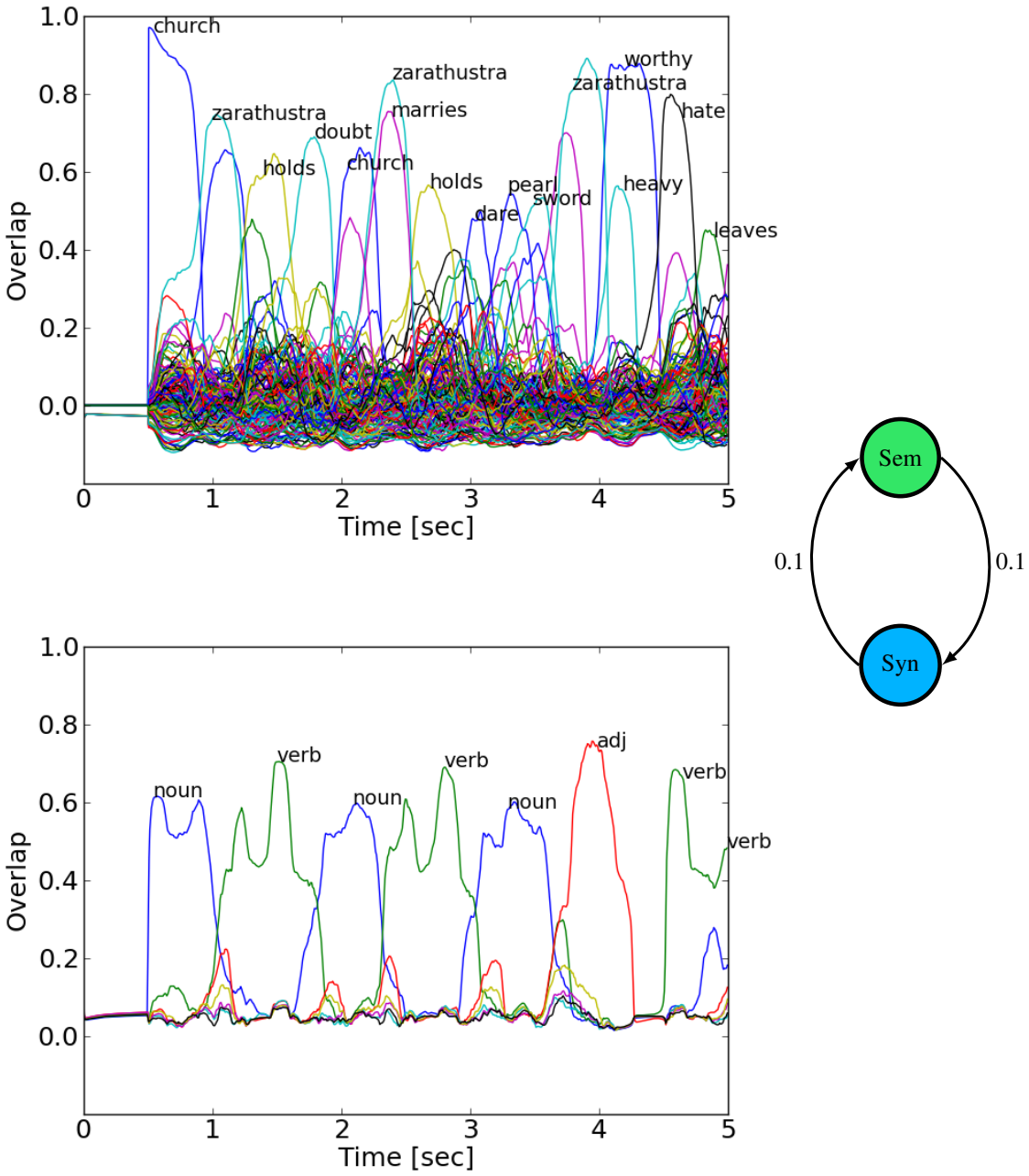


Figure 4-14: **Weak bidirectional interactions are too soft:** the interaction between the semantic sub-network (top) and the syntactic sub-network (bottom), with the weights $c_{sem2syn} = 0.1$ and $c_{syn2sem} = 0.1$. The sentences produced by the interaction of these two sub-networks are written on the top. The parameters were set at $w = 1.6$, $\beta = 5$, $U = 0.0$; for the semantic sub-network, $C_{sem} = 94$, $\tau_{1\,sem} = 10$, $\tau_{2\,sem} = 200$, $\tau_{3\,sem} = 10000$; for the syntactic sub-network, $C_{syn} = 68$, $\tau_{1\,syn} = 5$, $\tau_{2\,syn} = 100$, $\tau_{3\,syn} = 5000$, $c_{auto} = 1.0$, and $c_{hetero} = 0.0$.

church, sheep	marry, dare, gets	sword, stones	need, leaves, stands	horse, churches	kill, leaves	Zarathustra, sheeps
noun	verb	noun	verb	noun	verb	noun

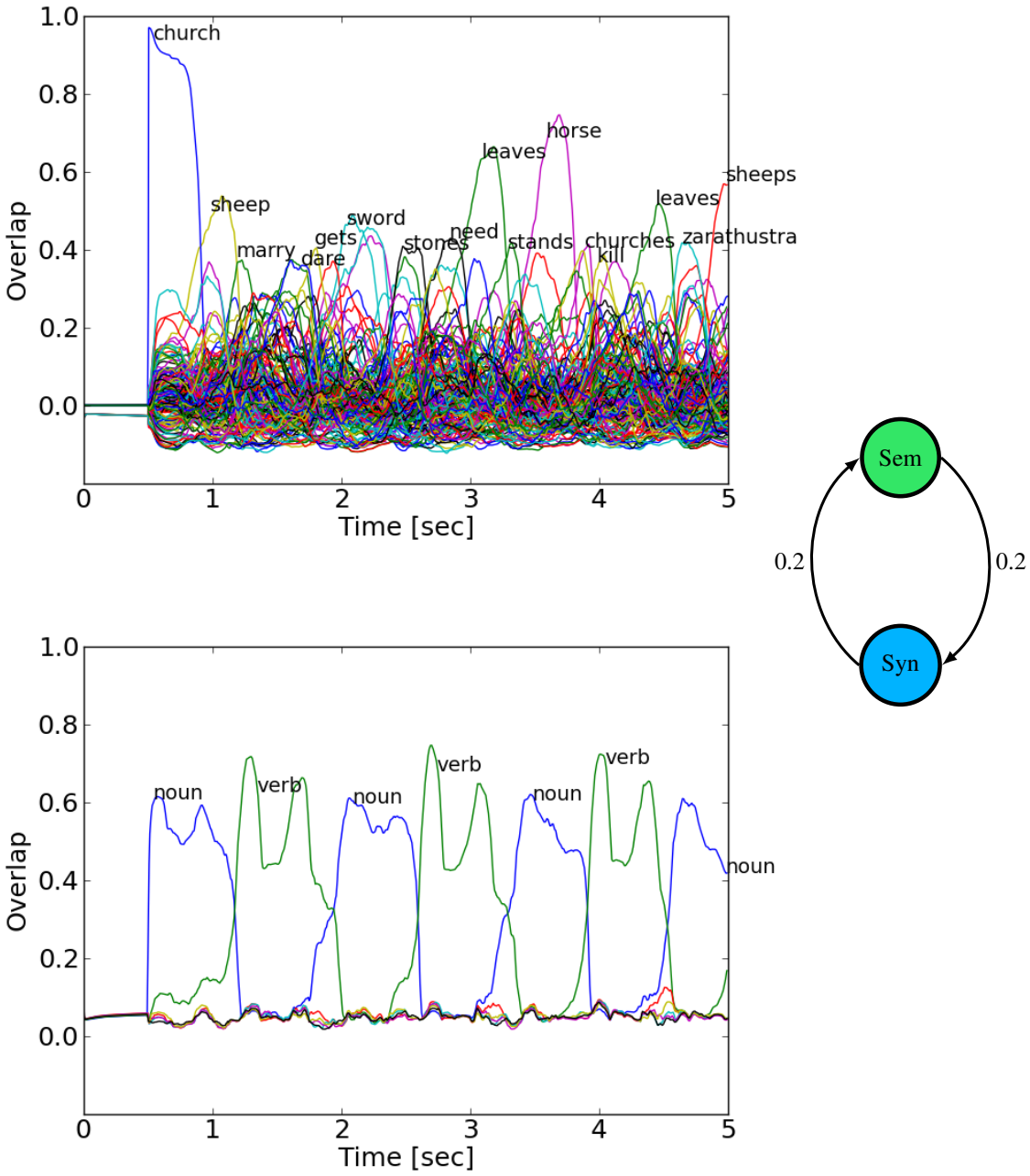


Figure 4-15: **Stronger interactions cause dynamics to stumble:** the interaction between the semantic sub-network (top) and the syntactic sub-network (bottom), with the weights $c_{sem2syn} = 0.2$ and $c_{syn2sem} = 0.2$. The sentences produced by the interaction of these two sub-networks are written on the top. The parameters were set at $w = 1.6$, $\beta = 5$, $U = 0.0$; for the semantic sub-network, $C_{sem} = 100$, $\tau_{1\,sem} = 10$, $\tau_{2\,sem} = 200$, $\tau_{3\,sem} = 10000$; for the syntactic sub-network, $C_{syn} = 75$, $\tau_{1\,syn} = 5$, $\tau_{2\,syn} = 100$, $\tau_{3\,syn} = 5000$, $c_{auto} = 1.0$, and $c_{hetero} = 0.0$.

4.5 Discussion

We investigated the behaviour of a Potts associative memory network that includes two components, semantic and syntactic, in which the words of the BLISS language were stored. First, we confirmed that the network parameters had the effects that we expected from a previous analysis (Russo and Treves, 2012). We observed that turning on the hetero-associative learning rule in the syntactic sub-network, when considered alone, produces more grammatically valid transitions between word categories. When investigating the interaction between the semantic and syntactic sub-networks, we observed that if the syntactic sub-network provides a cue for the retrieval of words in the semantic sub-network with randomly correlated patterns interesting changes in word transitions occur in the semantic sub-network, in some cases, producing more grammatically valid sequences of words.

Through the interaction between the semantic and syntactic components of the network, we approached in a novel way the problem of temporarily binding words in a sentence, without using algebraic operations (Stewart and Eliasmith, 2009) or highly complex connections between sentence constructs (Velde and de Kamps, 2006). If in the syntactic sub-network one of the syntactic categories (e.g. nouns) are retrieved—namely, the Potts units representing that category get activated—these active units serve as a cue for the retrieval of a word belonging to that category in the semantic sub-network. Among the words belonging to that category the semantic sub-network more probably retrieves the word that is semantically more correlated with preceding retrieved words.

Beside the binding problem, the problem of multiple instances needs to be addressed in a language model (Jackendoff, 2002); for instance, how could a language model distinguish multiple instances of the word *star* in ”The little star is beside the big star”? Our study provides some indication as to how this issue may be approached. Though in the Potts network there is only one representation for a word (e.g. *star*), stored as an attractor, the network approaches this attractor from different paths (either from *little* or from *big*); thus, network configurations for these two instances will not be identical. Moreover, since the activation of the Potts units is graded, the network might retrieve these two instances of the same word with different degrees of overlap.

Another problem that needs to be addressed by a language model is that of variable, related to reasoning processes (Jackendoff, 2002); for example, *John gives Mary a book* implies that *Mary owns a book*. At this early stage, the Potts network was only aimed at spontaneously producing sequences of encoded words. However, considering the high dimensionality of the Potts network, one could possibly add an additional reasoning component.

Though we successfully encoded the highly correlated patterns in the syntactic sub-network through appropriate normalization, we failed to store in the semantic sub-network the low correlated patterns generated and stored by the procedure explained in the previous chapter; when using those representations, the latches occurred mostly between the two patterns in an unusually highly correlated pair. How to store correlated patterns in a Potts network is a question that needs to be rigorously investigated in future studies.

Here we have demonstrated a few samples of the Potts network behaviour in its primitive stage of speaking; we have clearly not exhaustively explored the space of all possible cues—the starting word of a sequence—nor have we derived statistics about word sequences produced with different parameters in the network model or in the training language (the different semantic models in BLISS). These statistics should be compared with the information measures observed in the BLISS sentences. We note that within our limited experiments we have not yet observed the retrieval of some BLISS words, such as function words.

4.5.1 Future Directions

Although we distinguished the representation of singular and plural forms of the words in the syntactic sub-network, for simplicity we only demonstrated the overlap of the network configuration with the main lexical categories (i.e. nouns, verbs, ...), disregarding their singular and plural forms; this distinction should be added in future implementations. Further, the current implementation of the Potts network was not trained with end punctuation marks, thus there is no distinction between separate sentences; these marks could be included as additional attractors in the network.

Moreover, the network was trained on one language. A bilingual Potts network would require new word transitions belonging to the second language to be encoded in the syntactic sub-network; whether these transitions can be stored in the same syntactic sub-network or whether an additional syntactic component would be needed depends on the storage capacity of the network. Many aspects of the capacity question are still unknown, in particular for a network that uses a hetero-associative learning rule.

Speculative and based on evolutionary aspects, the large number of connections and the resulting large storage capacity of the human cortical network might have endowed it with an ability for latching beyond that of a mere associative memory network (Treves, 2005).

The stage of language acquisition and development is missing in the Potts network, where we have manually implemented words and stored the words on the network at the start of each simulation. Although the auto-associative learning rule stores each word independently of the others, the question of a self-organized acquisition of words and rules remains unaddressed by our model.

The question of whether the brain produces sequences of words in a manner similar to the Potts network is one question that needs to be investigated by experimental studies. Designing a careful experiment on spontaneous speech production might be difficult; However, one might design a speech perception task that examines, through brain signal analyses, the expectation of participants who are trained on the BLISS language upon listening to BLISS sentences.

In summary, we have implemented a Potts attractor neural network that stores the words of BLISS, the artificial language we constructed, and learns to some, admittedly limited, extent the transitions between words in its syntactic component, and produces sequences of words.

5

Conclusion

The main question of this Thesis is how we humans are able to produce sequences of words. To address this question, we have explored a novel approach that involves first constructing an artificial language of intermediate complexity and then implementing 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 inspired manner on two components: one semantic and one syntactic.

As the training language of the network, we have constructed BLISS, an artificial language of limited complexity that mimics natural languages by possessing a grammar of about 40 production rules, a vocabulary size of 150 words, and a definition of semantics, that is statistical dependences between words, beyond that due to production rules. The effect of introducing semantics with such a limited vocabulary was quantified using methods of information theory and found to be small, but still measurable.

As a sentence production model, we have implemented a Potts attractor neural network, whose units hypothetically represent patches of cortex. The choice of the Potts network, for sentence production, has been mainly motivated by the *latching* dynamics it exhibits; that is, an ability to spontaneously hop, or latch, across memory patterns, which have been stored as dynamical attractors, thus producing a long or even infinite sequence of patterns, at least in some regimes.

We have encoded words of BLISS, our artificial language, into the Potts network with two components, semantic and syntactic. We have also made a distinction between the

encoding of function words and content words. 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.

In order to generate the semantic and syntactic representation of words in a distributed fashion, we have used a generating algorithm that reflects the variable degrees of correlation between words. To encode a word representation into the Potts network, we have used an auto-associative learning rule—associating a word to itself—for the semantic sub-network; and we have used a hetero-associative—in addition to an auto-associative rule—for the syntactic sub-network, in order to associate different words together through the statistics of word transitions in BLISS.

Having stored the BLISS words, we investigated the behaviour of the Potts network. We first illustrated the influence of key parameters of the network by giving some examples of the latching transition using different values for these parameters. Next, we demonstrated the individual behaviour of the semantic and syntactic sub-networks. Finally, we considered the interaction between these two sub-networks, to see what word sequences the Potts network produces.

We observed that turning on the hetero-associative learning rule in the syntactic sub-network, when considered alone, produces more grammatically valid transitions between word categories. When investigating the interaction between the semantic and syntactic sub-networks, we observed that if the syntactic sub-network provides a cue for the retrieval of words in the semantic sub-network with randomly correlated patterns interesting changes in word transitions occur in the semantic sub-network, in some cases, producing more grammatically valid sequences of words.

Though promising, the talkative Potts network implemented here is at its preliminary stage. The capacity of the network for storing rules, correlated words, and several languages need to be investigated in future studies. The stage of language acquisition and development is missing in the Potts network, where we have manually implemented words and stored the words on the network at the start of each simulation. Despite these limitations, I hope this research opens a new path toward understanding mechanisms underlying

sentence production in the brain.

Bibliography

Abdollah-nia MF, Saeedghalati M, Abbassian A (2012) Optimal region of latching activity in an adaptive Potts model for networks of neurons. *Journal of Statistical Mechanics* 2012:P02018. 50

Abeles M (1991) *Corticonics: Neural Circuits of the Cerebral Cortex* Cambridge University Press. 40

Ahissar E, Vaadia E, Ahissar M, Bergman H, Arieli A, Abeles M et al. (1992) Dependence of cortical plasticity on correlated activity of single neurons and on behavioral context. *Science* 257:1412–1415. 40

Altmann G, Kamide Y (1999) Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition* 73:247–264. 30

Amit D, Gutfreund H, Sompolinsky H (1987) Information storage in neural networks with low levels of activity. *Physical Review A* 35:2293–2303. 46

Amit D (1992) *Modeling Brain Function: The World of Attractor Neural Networks* Cambridge University Press. 42

Bahlmann J, Schubotz RI, Friederici AD (2008) Hierarchical artificial grammar processing engages Broca's area. *Neuroimage* 42:525–534. 6, 30

Bicknell K, Elman J, Hare M, McRae K, Kutas M (2010) Effects of event knowledge in processing verbal arguments. *Journal of Memory and Language* 63:489–505. 30

Biran M, Friedmann N (2011) The representation of lexical-syntactic information: Evidence from syntactic and lexical retrieval impairments in aphasia. *Cortex* 48:1103–1127. 41

Borensztajn G (2011) 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 Amsterdam. 2

Braitenberg V, Schüz A (1991) *Anatomy of the Cortex: Statistics and Geometry.* Springer-Verlag Publishing. 40, 42, 43

Broca P (1861) Perte de la parole, ramollissement chronique et destruction partielle du lobe antérieur gauche du cerveau. *Bull Soc Anthropol* 2:235–238. 39

Caramazza a, Shelton JR (1998) Domain-specific knowledge systems in the brain the animate-inanimate distinction. *Journal of Cognitive Neuroscience* 10:1–34. 41

Charniak E (1995) Parsing with context-free grammars and word statistics. *Brown University* pp. 598–603. 9

Charniak E (2000) A maximum-entropy-inspired parser In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, pp. 132–139. 30

Chomsky N (1957) *Syntactic Structures* Mouton, The Hague. 29

Chomsky N (1959) On certain formal properties of grammars. *Information Control* 2:137–167. 9

Christiansen MH (2000) Using artificial language learning to study language evolution: Exploring the emergence of word order universals In *The Evolution of Language: 3rd International Conference*, pp. 45–48. 5

Cover T, Thomas J (1991) *Elements of Information Theory* Wiley Online Library. 20

Crescentini C, Shallice T, Macaluso E (2010) Item retrieval and competition in noun and verb generation: An fmri study. *Journal of Cognitive Neuroscience* 22:1140–57. 69

Damasio H, Grabowski T, Tranel D, Hichwa R, Damasio A et al. (1996) A neural basis for lexical retrieval. *Nature* 380:499–505. 41

de Diego-Balaguer R, Fuentemilla L, Rodriguez-Fornells A (2011) Brain dynamics sustaining rapid rule extraction from speech. *Journal of Cognitive Neuroscience* 23:3105–3120. 6, 30

Deacon T (1989) Holism and associationism in neuropsychology: An anatomical synthesis. *Integrating Theory and Practice in Clinical Neuropsychology* pp. 1–47. 39

Deacon T (1992) Cortical connections of the inferior arcuate sulcus cortex in the macaque brain. *Brain Research* 573:8–26. 40

Ehrlich S, Rayner K (1981) Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior* 20:641–655. 30

Elman JL (1993) Learning and development in neural networks: the importance of starting small. *Cognition* 48:71–99. 2, 70

Elman JL (1991) Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning* 7:195–225. 2

Frank SL, Bod R (2011) Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science* 22:829–34. 41

Friederici aD, Opitz B, von Cramon DY (2000) Segregating semantic and syntactic aspects of processing in the human brain: an fMRI investigation of different word types. *Cerebral Cortex* 10:698–705. 40, 69

Friederici A, Schoenle P (1980) Computational dissociation of two vocabulary types: Evidence from aphasia. *Neuropsychologia* 18:11–20. 40, 69

Friederici AD, Steinhauer K, Pfeifer E (2002) Brain signatures of artificial language processing: Evidence challenging the critical period hypothesis. *Proceedings of the National Academy of Sciences of the United States of America* 99:529–534. 5, 6

Friedmann N (2001) Agrammatism and the psychological reality of the syntactic tree. *Journal of Psycholinguistic Research* 30:71–90. 41

Fulvi Mari C, Treves A (1998) Modeling neocortical areas with a modular neural network. *Biosystems* 48:47–55. 43, 46

Gayler RW (2003) Vector symbolic architectures answer Jackendoff's challenges for cognitive neuroscience In *The Joint International Conference on Cognitive Science*. 2, 70

Gomez R (2000) Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences* 4:178–186. 5, 6

Grodzinsky Y (1984) The syntactic characterization of agrammatism. *Cognition* 16:99–120. 41

Gruening A (2004) Neural networks and the complexity of languages Ph.d. thesis, School of Mathematics and Computer Science, University of Leipzig. 33

Hale J (2001) A probabilistic Earley parser as a psycholinguistic model In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pp. 1–8. 31

Hebb D (1949) *The Organization of Behavior: A Neuropsychological Theory* Wiley. 40, 46

Herrmann M, Ruppin E, Usher M (1993) A neural model of the dynamic activation of memory. *Biological Cybernetics* 68:455–463. 50

Hochmann JR, Endress AD, Mehler J (2010) Word frequency as a cue for identifying function words in infancy. *Cognition* 115:444–457. 6, 30

Hopfield J (1982) Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences* 79:2554–2558. 42, 50

Horn D, Usher M (1989) Neural networks with dynamical thresholds. *Physical Review A* 40:1036–1044. 49

Hummel J, Holyoak K (1997) Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review* 104:427–466. 1, 70

Hummel JE, Holyoak KJ (2003) A symbolic-connectionist theory of relational inference and generalization. *Psychological Review* 110:220–264. 1

Huyck CR (2009) A psycholinguistic model of natural language parsing implemented in simulated neurons. *Cognitive Neurodynamics* 3:317–330. 1

Jackendoff R (1999) The representational structures of the language faculty and their interactions. *The Neurocognition of Language* pp. 37–79. 40

Jackendoff R (2002) *Foundations of Language: Brain, Meaning, Grammar, Evolution* Oxford University Press, USA. 95, 96

Jurafsky D, Martin JH (2000) *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition* Prentice Hall. 8, 9, 31

Kanter I (1988) Potts-glass models of neural networks. *Physical Review A* 37:2739–2742. 43, 47

Kinder A, Lotz A (2009) Connectionist models of artificial grammar learning: What type of knowledge is acquired? *Psychological Research* 73:659–673. 5

Klein D, Manning C (2003) Accurate unlexicalized parsing In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pp. 423–430. 30

Knowlton BJ, Squire LR (1996) Artificial grammar learning depends on implicit acquisition of both abstract and exemplar-specific information. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22:169–181. 6

Kohl MM, Paulsen O (2010) The roles of GABAB receptors in cortical network activity. *Advances in Pharmacology* 58:205–29. 49

Kropff E, Treves A (2005) The storage capacity of Potts models for semantic memory retrieval. *Journal of Statistical Mechanics: Theory and Experiment* 2005:P08010. 46, 47, 70

Kropff E, Treves A (2006) The complexity of latching transitions in large scale cortical networks. *Natural Computing* 6:169–185. 3, 43, 50, 52

Kutas M, Hillyard SA (1984) Brain potentials during reading reflect word expectancy and semantic association. *Nature* 307:161–163. 30

Lany J, Saffran J (2010) From statistics to meaning. *Psychological Science* 21:284–291. 7

Lerner I, Bentin S, Shriki O (2012) Spreading activation in an attractor network with latching dynamics: Automatic semantic priming revisited. *Cognitive Science* 36:1339–1382. 69

Levy R (2008) Expectation-based syntactic comprehension. *Cognition* 106:1126–77. 31

Longobardi G, Guardiano C (2009) Evidence for syntax as a signal of historical relatedness. *Lingua* 119:1679–1706. 33

MacKay D (2003) *Information Theory, Inference, and Learning algorithms* Cambridge University Press. 21

Mahon BZ, Caramazza A (2009) Concepts and categories: A cognitive neuropsychological perspective. *Annual Review of Psychology* 60:27–51. 41

Manning CD, Raghavan P, Schütze H (2008) *Introduction to Information Retrieval* Cambridge University Press. 31

Manning CD, Schuetze H (1999) *Foundations of Statistical Natural Language Processing* The MIT Press. 6, 8

Marcus G, Vijayan S, Bandi Rao S, Vishton P (1999) Rule learning by seven-month-old infants. *Science* 283:77–80. 6

McRae K, de Sa VR, Seidenberg MS (1997) On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology* 126:99–130. 42

McRae K, Cree GS, Seidenberg MS, Mcnorgan C (2005) Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods* 37:547–559. 11, 54

Mitchell TM, Shinkareva SV, Carlson A, Chang KM, Malave VL, Mason Ra, Just MA (2008) Predicting human brain activity associated with the meanings of nouns. *Science* 320:1191–5. 42, 53, 69

Monaghan P, Chater N, Christiansen M (2005) The differential role of phonological and distributional cues in grammatical categorisation. *Cognition* 96:143–182. 32

Monaghan P, Christiansen M, Chater N (2007) The phonological-distributional coherence hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive Psychology* 55:259–305. 32

Mueller J, Oberecker R, Friederici A (2009) Syntactic learning by mere exposure–An ERP study in adult learners. *BMC Neuroscience* 10:89. 5

O'Kane D, Treves A (1992) Short- and long-range connections in autoassociative memory. *Journal of Physics A: Mathematics and General* 25:5055–5069. 42, 43

Opitz B, Friederici AD (2003) Interactions of the hippocampal system and the prefrontal cortex in learning language-like rules. *Neuroimage* 19:1730–1737. 6

Pallier C, Devauchelle AD, Dehaene S (2011) Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences of the United States of America* 108:2522–7. 41

Pena M, Bonatti LL, Nespor M, Mehler J (2002) Signal-driven computations in speech processing. *Science* 298:604–607. 5

Petersson KM, Folia V, Hagoort P (2010) What artificial grammar learning reveals about the neurobiology of syntax. *Brain and Language* 120:83–95. 5, 40

Piantadosi ST, Tily H, Gibson E (2011) Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences of the United States of America* 108:3526–3529. 31

Pirmoradian S, Treves A (2011) BLISS: an artificial language for learnability studies. *Cognitive Computation* 3:539–554. 2

Plate T (1991) Holographic reduced representations: Convolution algebra for compositional distributed representations In *Proceedings of the 12th International Joint Conference on Artificial Intelligence*, pp. 30–35. 2

Pollock J (1989) Verb movement, universal grammar, and the structure of ip. *Linguistic Inquiry* 20:365–424. 41

Pulvermüller F (1999) Words in the brain's language. *Behavioral and Brain Sciences* 22:253–336. 40

Reber A (1967) Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior* 6:855–863. 6

Richardson FM, Thomas MSC, Price CJ (2010) Neuronal activation for semantically reversible sentences. *Journal of Cognitive Neuroscience* 22:1283–1298. 33

Rosenkrantz D, Lewis P (1970) Deterministic left corner parsing In *Switching and Automata Theory, IEEE Conference Record of 11th Annual Symposium*, pp. 139–152. 9

Russo E, Namboodiri V, Treves A, Kropff E (2008) Free association transitions in models of cortical latching dynamics. *New Journal of Physics* 10:015008. 50, 52

Russo E, Pirmoradian S, Treves A (2010) Associative latching dynamics vs. syntax In *Advances in Cognitive Neurodynamics (II): Proceedings of the Second International Conference on Cognitive Neurodynamics*, p. 111. 70

Russo E, Treves A (2012) Cortical free-association dynamics: Distinct phases of a latching network. *Physical Review E* 85:1–18. 3, 48, 50, 51, 58, 70, 73, 95

Saffran JR, Aslin RN, Newport EL (1996) Statistical learning by 8-month-old infants. *Science* 274:1926–1928. 5

Shallice T (1988) Specialisation within the semantic system. *Cognitive Neuropsychology* 5:133–142. 41

Shallice T, Cooper R (2011) *The Organisation of Mind* OUP Oxford. 40, 69

Shapiro K, Caramazza A (2004) The organization of lexical knowledge in the brain: The grammatical dimension In Gazzaniga M, editor, *Cognitive neurosciences*, pp. 803–814. Cambridge, MA: MIT Press, 3rd edition. 40, 69

Shapiro K, Caramazza A (2003) The representation of grammatical categories in the brain. *Trends in Cognitive Sciences* 7:201–206. 40

Shapiro K, Shelton J, Caramazza A, Shapiro K, Shelton J, Caramazza A (2000) Grammatical class in lexical production and morhpological processing : Evidence from a case of fluent aphasia. *Cognitive Neuropsychology* 17:665–682. 40

Shieber S (1985) Evidence against the context-freeness of natural language. *Linguistics and Philosophy* 8:333–343. 29

Shukla M, White K, Aslin R (2011) Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. *Proceedings of the National Academy of Sciences of the United States of America* 108:6038–6043. 32

Smolensky P (1990) Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence* 46:159–216. 2

Sompolinsky H, Kanter I (1986) Temporal association in asymmetric neural networks. *Physical Review Letters* 57:2861–2864. 46, 48

Stewart T, Eliasmith C (2009) Compositionality and biologically plausible models In Werning M, Hinzen W, Edouard M, editors, *The Oxford handbook of compositionality*. Oxford University Press. 1, 2, 95

Toro J, Nespor M, Mehler J, Bonatti L (2008) Finding words and rules in a speech stream. *Psychological Science* 19:137–144. 32

Treves A (2005) Frontal latching networks: a possible neural basis for infinite recursion. *Cognitive Neuropsychology* 22:276–91. 43, 49, 50, 52, 61, 97

Tsodyks M, Feigelman (1988) The enhanced storage capacity in neural networks with low activity level. *Europhysics Letters* 6:101–105. 46

Tsodyks M, Markram H (1997) The neural code between neocortical pyramidal neurons depends. *Proceedings of the National Academy of Sciences of the United States of America* 94:719–723. 49

Ullman M, Pancheva R, Love T, Yee E, Swinney D, Hickok G (2005) Neural correlates of lexicon and grammar: Evidence from the production, reading, and judgment of inflection in aphasia. *Brain Language* 93:185–238. 6

Vaadia E, Haalman I, Abeles M, Bergman H, Prut Y, Slovin H, Aertsen A et al. (1995) Dynamics of neuronal interactions in monkey cortex in relation to behavioural events. *Nature* 373:515–518. 40

van Dam WO, van Dijk M, Bekkering H, Rueschemeyer SA (2012) Flexibility in embodied lexical-semantic representations. *Human Brain Mapping* 33:2322–33. 42

Velde FVD, de Kamps M (2006) Neural blackboard architectures of combinatorial structures in cognition. *Behavioral and Brain Sciences* 29:37–108. 1, 2, 70, 95

Vigliocco G, Vinson DP, Druks J, Barber H, Cappa SF (2011) Nouns and verbs in the brain: A review of behavioural, electrophysiological, neuropsychological and imaging studies. *Neuroscience and Biobehavioral Reviews* 35:407–26. 41

Warrington EK, Shallice T (1984) Category specific semantic impairments. *Brain* 107:829–853. 41

Wernicke C (1874) *Der Aphasische Symptomencomplex: Eine Psychologische Studie auf Anatomischer Basis* Cohn & Weigert. 39