Neurobiology of incremental speech comprehension

Hun Seok Choi

Clare College

University of Cambridge

September 2018

This dissertation is submitted for the degree of Doctor of Philosophy

Acknowledgement

I would like to thank my supervisor Professor Lorraine Tyler for the advice during the course of my studies. Special thanks also to Dr. Barry Devereux and Dr. Billi Randall for their help and guidance.

I express my gratitude to Dr. Andrew Thwaites, Dr. Alex Clark, Dr. Bingjiang Lyu, Dr. Cai Wingfield, Dr. Darren Price and Professor William Marslen-Wilson for various helps and useful discussions.

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text

It does not exceed the prescribed word limit for the relevant Degree Committee.

Neurobiology of incremental speech comprehension

Hun S. Choi

Abstract

Understanding spoken language requires the rapid transition from perceptual processing of the auditory input through a variety of cognitive processes involved in constructing the mental representation of the message that the speaker is intending to convey. Listeners carry out these complex processes very rapidly and accurately as they hear each word incrementally unfolding in a sentence. However, little is known about the specific spatiotemporal patterning of this wide range of incremental processing operations that underpin the dynamic transitions from the speech input to the development of a meaning interpretation of an utterance. This thesis aims to address this set of issues by investigating the spatiotemporal dynamics of brain activity as spoken sentences unfold over time in order to illuminate the neurocomputational properties of the human language processing system and determine how the representation of a spoken sentence develops incrementally as each upcoming word is heard.

Using a novel application of multidimensional probabilistic modelling combined with models from computational linguistics, I developed models of a variety of computational processes associated with accessing and processing the syntactic and semantic properties of sentences and tested these models at various points as sentences unfolded over time. Since a wide range of incremental processes occur very rapidly during speech comprehension, it is crucial to keep track of the temporal dynamics of the neural computations involved. To do this, I used combined electroencephalography and magnetoencephalography (EMEG) to record neural activity with millisecond resolution and analyzed the recordings in source space using univariate and/or multivariate approaches. The results confirm the value of this combination of methods in examining the properties of incremental speech processing. My findings corroborate the predictive nature of human speech comprehension and demonstrate that the effects of early semantic constraint are not dependent on explicit syntactic knowledge.

Contents

Chapter 1: Introduction 11

1.1. Theories of grammar and language comprehension 13

1.2. Constraints and prediction 14

1.3. Predictive computations involved in the multi-level speech processing: theoretical reviews 18

1.3.1. Modular theory and “syntax-first” 21

1.3.2. Parallel-interaction and constraint satisfaction approaches 29

1.4. Neuroanatomy of speech processing 35

1.4.1. Syntactic and semantic processing in the brain 36

1.4.2. Anatomical connectivity within language networks 40

1.4.3. Neurobiological models of speech comprehension 42

1.4.4. The lexicalist approach 45

1.5. Issues addressed in this thesis 46

1.6. Preparing EEG/MEG data for the investigation of speech comprehension 48

1.6.1. Electro- and Magneto-encephalography 48

1.6.2. Participants 50

1.6.3. Stimuli and Procedure 51

1.6.4. EMEG pre-processing 52

1.6.5. EMEG source reconstruction 53

Chapter 2: Computational modelling of the incremental processing of a sentence 57

2.1. Bayesian Belief Updating (BBU) 58

2.2. Computational and behavioural modelling of prediction 62

2.3. Modelling prediction with neural networks 64

2.3.1. Capturing non-linear patterns in the data using non-linear functions of linear classification algorithms 65

2.3.2. Output layer and softmax 68

2.3.3. Further implementation details 68

2.3.4. Adding recurrence in the network 70

2.3.5 Incremental language processing in a LSTM neural network 73

2.4. Quantifying the “degree” in prediction: the information-theoretic framework 76

2.5. Constraints modelling 79

2.5.1 Modelling a constraint on syntax 79

2.5.2. Modelling constraints on semantics 81

Chapter 3: Decoding the real-time neurobiological properties of incremental speech comprehension 93

3.1. Overview 95

3.2. Decoding the multivariate pattern of neural activity 97

3.2.1. Representational dissimilarity matrix (RDM) and distance metrics 97

3.2.2. Cluster statistics 102

3.2.3. Multiple comparisons correction 104

3.3. Models of constraints 105

3.4. Additional analysis: activation of the generalized semantic contents of a constraining word in a sentence 109

3.5. Results 110

3.5.1(a) Incremental representational constraints: Semantic 114

3.5.1(b) Incremental representational constraints: Syntactic 116

3.5.2(a) Integration effects: Semantic 118

3.5.2(b) Integration effects: Syntactic 119

3.5.3. Activation of the semantic contents 120

3.6. Discussion 123

3.6.1. Early RH computation of incremental ‘event-structure’ representations 123

3.6.2. LH computation of bottom-up lexically-driven syntactic constraints 125

3.6.3. WordNet-MDL based approach of modelling semantic constraints 127

3.6.4. Conclusion 128

Chapter 4: Decoding the internal representation of a predictive machine and testing its relation with the neural computations 131

4.1. Sequential vs. hierarchical processing accounts of human speech comprehension 132

4.2. Connectionist models in a parallel distributed processing (PDP) framework 135

4.3. Decoding the pattern of activation in the LSTM internal and output layers 138

4.3.1. Sanity checks and methods 138

4.3.2. Results 140

4.3.3. Summary and discussion: Comparing LSTM with human speech processing 143

4.4. Neurocognitive hypotheses: characterization of the neural response patterns using the LSTM layers 144

4.4.1. Implications of the section 4.3 on neurocognitive hypotheses 145

4.4.2 Epochs and analysis 146

4.5. Results 148

4.5.1. (i) Subject Noun and Verb (Epoch V): 149

4.5.2. (ii) Complement onset (Epoch CFW): 156

4.5.3. (iii) Complement noun onset (Epoch CN1): 156

4.5.4 (iv) LSTM models vs. Behavioural and corpus-based models 157

4.5.5. (v) Additional analysis (Epoch CN2): 161

4.5.6. (vi) Summary 163

4.6. Discussion 164

4.6.1. Activating lexical properties of an input word in a sentence 165

4.6.2. Incremental constraint and prediction 167

4.6.3. Integration 170

4.6.4. Methodology: insights on the different modelling aspects 171

Chapter 5: General discussion 175

5.1. Advantages of computational modelling in explaining neurobiological data 175

5.2. Predictive processing in incremental speech comprehension 177

5.3. Neurobiological account of syntax and semantics in predictive computations for incremental speech comprehension 182

5.4. Conclusion 186

References 187

Appendices 203

Appendix 1: Merge and integration 203

Appendix 2: A list of all sentence stimuli 207

Appendix 3: Weber-Fechner's Law 212

Appendix 4: derivation of the LDA training algorithm (collapsed Gibbs sampler) 214

Appendix 5: Representing the constraint of each verb on its argument through an optimized set of "synsets" (i.e. conceptual senses) 218

Appendix 6: Derivatives of the non-linear functions 230

Appendix 7: Backpropagation and gradient learning 233

Chapter 1: Introduction

Speech comprehension engages complex cognitive processes, including the rapid activation of the lexical properties of incrementally unfolding words and their on-line integration into the developing sentence. However, listeners can readily process every word and easily interpret it in the context in which it is heard. To do this, listeners engage a number of complex processes over a short period of time including acoustic analysis of the speech waveform, its mapping onto phonemic and lexical level representations, retrieving syntactic and semantic properties associated with the lexical object, updating the internal representation of the message with respect to the retrieved lexical information and using the updated representation to constrain the upcoming words. It is now widely acknowledged that the ability to integrate the available information from the context in order to facilitate the processing of the bottom-up inputs provides a basis for such complex processes. This thesis aims to address the wide range of incremental processing operations that underpin such rapid and efficient understanding of speech and illuminate the properties of the neurobiological system in which they are instantiated. In particular, I investigated the neural computations involved in constraining and guiding the interpretation of upcoming words in sentences, enabling listeners to rapidly integrate each word into the developing sentential context. Such a dynamic process of constraining and integrating incrementally unfolding information is a crucial part of understanding speech comprehension yet is often overlooked in neurobiological models of speech comprehension in the literature.

Addressing this issue requires clarifying the nature of linguistic computations during incremental speech comprehension. This dissertation focuses on the following set of questions that have either been controversial or not been thoroughly investigated in the literature: 1) What are the linguistic bases of predictive computations? 2) Are the syntactic properties of constraints activated prior to the activation of semantic properties? 3) Do listeners utilize these constraints to guide their interpretation? 4) To what extent is human speech comprehension incremental? (or, more specifically, do these predictive computations occur word-by-word in a sentence?) and 5) Is it possible for a model, that learns statistical relations among different words through a large corpus but does not have any explicit knowledge of syntax, to explain predictive processing in human speech comprehension? These questions are addressed from a neurobiological perspective by characterizing the encoded information in the spatiotemporal dynamics of neural activity during natural speech comprehension.

Using a novel application of multidimensional probabilistic modelling, I developed models of a variety of computational processes associated with the syntactic and semantic properties of words. This approach provides informative and realistic models of incremental computations in terms of how listeners experience language. This approach is particularly well-suited to address the aforementioned questions because they characterize a variety of linguistic properties from multi-level constraints in the form of a distribution. I varied the extent of the context on which syntactic and semantic constraints are based using behavioural models from pre-test data (the full-context models) and computational models using corpus data (single word context models). The combination of these behavioural and corpus-based approaches enables us to construct models of constraint based on the entire preceding context while preserving the accurate statistical information associated with every predicted word.

Moreover, I also used a sophisticated connectionist model trained on the corpus data to model the way that each word is processed in an optimized predictive machine. This connectionist model is in the form of LSTM (long, short-term memory) neural network (Jozefowicz, Vinyals, Schuster et al., 2016) that captures incremental processing of every word through recurrent connections and how it changes the internal state without guidance from syntactic knowledge. I explored the explanatory value of these models of contextual (context-based) and lexical (single-word-based) constraints during speech comprehension to address the neurocomputational questions above.

Since a wide range of incremental processes occur very rapidly during speech comprehension, it is crucial to keep track of the temporal dynamics of the neural computations involved. To do this, I used electroencephalography and magnetoencephalography (EMEG) to record neural activity with millisecond resolution. These recordings were analyzed in the source space using multivariate approaches given the neural activity that inherently varies across space (vertices) and time. In particular, a variant of an MVPA (multi-voxel pattern analysis) approach, known as representational similarity analysis (RSA; Kriegeskorte, Mur & Bandettini, 2008), was used which is well-suited to investigating the neurocognitive processes through characterizing the information encoded in the multivariate patterns of neural activity using the multidimensional (distributional) models of constraints. Using these modelling and analysis approaches, I aimed to address the aforementioned questions and to elucidate the way in which the complex predictive processes are neurobiologically instantiated throughout this thesis.

1.1. Theories of grammar and language comprehension

Understanding a word in solitary use does not require grammar. Grammar is the structure of language that guides comprehenders to interpret a word in context of the other words. Therefore, utilizing a set of combinatiorial rules, which we refer to as grammar, is what allows humans to communicate a highly complex message that consists of more than one linguistic unit. Researchers have sketched different maps of grammar based on different of architectural features and proposed a number of theories that explains such combinatorial operations during language comprehension with different claims. In this section, I briefly describe three major grammar theories built upon different assumptions to provide theoretical motivations to the psycholinguistic theories and their hypotheses regarding the incremental speech comprehension in the next section.

Generative grammar is one of the well-known theories introduced and developed by the influential linguist, Noam Chomsky (Chomsky, 1964, 1981, 1982, 1993). Its basic architecture contains three levels of representation (i.e. syntactic, semantic, and phonological) where phonological and semantic components are purely interpretive (Chomsky, 1964). Rather, only syntactic component is computational such that there are only two possible mapping processes between these linguistic levels: syntax to semantics (D-structure) and syntax to phonology (S-structure). What enable such mapping are a set of rules that relate each linguistic unit to each other (constituency relation) and a set of operations such as "merge" (see Appendix 1) and "move" (i.e. an operation that allows the movement of constituents to overcome the discontinuity or displacement in constituency grammar). Therefore, this theory is strongly derivational and unidirectional originating from syntax.

As opposed to the framework of generative grammar, the parallel architecture framework developed by Ray Jackendoff (1997) defines syntax, semantics and phonology as three independent components with its own symbolic primitives and principles of combination. In this framework, there is no one-to-one mapping between syntax and the other levels but such inter-level relation is rather licensed by "interface constraints". Hence, each autonomous structure of a particular component is licensed by its unique internal constraints as well as bidirectional interfaces to the other levels. Consequently, this view argues against any theories built upon a syntactocentric derivation that puts syntax in its ruling position ahead of phonology or semantics and the sequentially-ordered derivations of each constituent are replaced by "parallel constraint checking" emanating from the autonomous structures.

The last grammar theory considered here is a specific development of functional grammar known as functional discourse grammar (Hengeveld, 2004; Hengeveld & Mackenzie, 2008, 2010). As communication in a natural language environment almost always requires interpersonal interactions, this theory introduces four levels of representation that are hierarchically organized in a following order (top-down): pragmatic, semantic, morphosyntactic and phonological. As opposed to syntactocentric derivational theories claiming that computation always starts from syntax, functional discourse grammar rather proposes that the pragmatics/semantics is where computation starts from which is, then, translated into the formal level of representation (morphosyntactic/phonological). This is a clear example of top-down (unidirectional) pragmato-semantocentric grammar that emphasises pragmatics/semantics influence over syntax (non-derivational).

The architectural features upon which these grammar theories are built provide theoretical grounds to different psychological theories of speech comprehension.

1.2. Constraints and prediction

To be clear about the usage of terms throughout this thesis, I define the term "prediction" as the influence of prior beliefs on the state of the language processing system before the listener hears the bottom-up input. The term "constraints" refers to the prior beliefs themselves. Hence, I define "constraints" as information that can predictively alter the state of the human language system. The benefits/costs of making prediction depend on whether the continuation turns out to be as expected. However, it is now widely acknowledged that prediction brings facilitatory effects to fast and accurate speech comprehension in the noisy and ambiguous natural language environment (Kuperberg & Jaeger, 2016) and researchers have already found evidence for predictive processes during speech comprehension (see below). In my view, what determines the usefulness of prediction is the amount of information in the context and it has a direct implication on the level (degree of specificity) of prediction:

a) The day was breezy so the boy went outside to fly a ...
b) Flying ...

In a), the context provides rich information towards a specific word "kite", whereas, in b), the context vaguely prefers an object or a subject that can fly. Hence, the "amount of *information" that guides the level of prediction is likely reflected by its entropy (see Chapter 2 for more details). The probabilistic models that I discuss and develop throughout this thesis provides a statistically optimized constraint as a probability distribution across abstracted candidates and explains the way to quantify the degree of mismatch (incorrectness) between a predicted and an actual continuation as well as its psycholinguistic implications (see Chapter 2).*

Also, throughout this thesis, the term "target" is used to refer any linguistic units that are predicted by a preceding "context". It can be a word, a phrase or a sentence at the lexical, the semantic or the syntactic level. When modelling the changing beliefs during incremental speech processing (see Chapter 2), the target can be a unit that is being predicted if it has not been revealed yet or an input that is being integrated if it is being revealed. Similarly, the term "input" throughout this thesis is used to refer to any linguistic units including a word, a phrase or a sentence that has been or is being revealed at the current time. These generic terms are used to describe the conceptual framework of predictive processing in which multiple linguistic units at multiple levels can be constrained.

Language comprehension involves interactive processes of constraining the upcoming input and analyzing it at different linguistic levels (Kuperberg, 2016). This can be seen from well-established linguistic phenomena such as garden-path effects or ambiguity resolution which cannot be explained unless listeners utilize the context to facilitate the processing of an utterance. Ambiguity is a natural property of language which renders the linguistic input to be interpreted in different ways with respect to the context. For example, at least 80% of English words have more than one dictionary definition which makes them semantically ambiguous (e.g. "*blind*" in "*The blind on the window kept out the sun*"). The syntactic interpretation of a word can also be ambiguous: in the sentence "*The developer knew that building services are supplied by the local council*", the word "*building*" can be interpreted either as a subject itself in a gerundive phrase or as a modifier of the following noun "*services*" until the disambiguating verb "*are*" appears in the sentence. Garden-path sentences refer to syntactically ambiguous sentences that have a dominant and a subordinate interpretation and mislead the parser to an incorrect (dominant) interpretation (Bever, 1970): for example, in the sentence "*The horse raced past the barn fell*", the parser is tricked to interpret the verb "*raced*" as a main verb in a sentence but it turns out to be a clausal verb in a reduced relative clause as "*fell*" is heard. If listeners wait until the sentence is fully heard, such temporary ambiguities will not affect their comprehension process at all. However, researchers have consistently found significant effects of such temporary ambiguities in human language comprehension (Rodd et al., 2005, 2010). Evidence from the psycho- and neuro-linguistic literatures converges to the claim that listeners actively constrain the likely continuations using the information provided by the context and reanalyse if the interpretation turns out to be incorrect (e.g. the way that the noun "*services*" is interpreted heavily depends on the preceding verb's ("*build*") preference for a direct object frame; Tyler & Marslen-Wilson, 1977; Tyler et al., 2013).

However, the degree to which and the way in which the human language system constrains the incrementally unfolding sentence is controversial. For example, the constraint-satisfaction theory (MacDonald, 1994) predicts that the system constrains the likely candidates simultaneously and refines them until only one candidate remains (Mellish, 1981) whereas the syntax-first theory (Frazier & Fodor, 1978) predicts that the system constructs the initial parse towards the simplest structure which can be updated and changed in a serial manner as information unfolds over time. In such a serial processing view, if the bottom-up input turns out to be incongruent with the parser's interpretation, the context (in conjunction with the disconfirming bottom-up input) should be fully reanalysed to provide the correct interpretation. However, given the massive number of possible continuations for any given context in a natural language environment, such serial processing is highly resource-demanding: *"why bother predicting just one candidate, only to be proven wrong?"* (Kuperberg & Jaeger, 2016, p. 34). Therefore, under the view that human language processing is predictive (see Delong, Urbach & Kutas., 2005; Federmeier, Wlotko, De Ochoa-Dewald & Kutas, 2007; Altmann & Mirkovic, 2009; Kuperberg & Jaeger, 2016), it is implicitly assumed that the parser predicts the likely candidates simultaneously in relation to one another.

Modelling prediction

One of the simplest ways of modelling such predictive processes is to compute a conditional probability distribution consisting of the likely candidates constrained by the prior context with a varying degree of expectancy (e.g. an N-gram model). Under the probability rule that $P(x_t|x_{t-1}, \ldots, x_{t-N}) = \frac{P(x_t, x_{t-1}, \ldots, x_{t-N})}{P(x_{t-1}, \ldots, x_{t-N})}$ where $P(x_t|x_{t-1}, \ldots, x_{t-N})$ represents the probability of an event x_t given that the series of events from x_{t-1} to x_{t-N} has occurred, this can simply be implemented by counting the co-occurrence frequency between the preceding context that contains N number of words and each of the candidates in a large corpus, and normalizing by the summed frequency across all candidates for a given context. In practice, the output probability values associated with each pair can be divided by the probability of the context in the pair calculated from the same corpus so that the probability of every candidate is not biased by the frequency of the preceding context[1]. This type of corpus-based modelling tends to be remarkably accurate in modelling the constraints at the individual word (lexical) level, assuming that the corpus contains sufficient samples to cover the entire distribution (see Jurafsky & Martin, 2009). However, as the size of context grows from a word to a phrase and from a phrase to a sentence, the required amount of data in the corpus also grows exponentially with each word added to the context. Therefore, it becomes practically difficult to model the constraints generated by a context consisting of more than three words. Such N-gram type model is neurobiologically implausible as well since the entire context is reduced to previous n-1 adjacent words (Frank et al., 2015).

Modelling brain activity based directly on the raw frequency of a word or words in a corpus implicitly assumes the brain as a large "lexicon" like the corpus from which the specific lexical knowledge associated with a word (or a set of words) is retrieved. However, a different view suggests "words are not mental objects that reside in a mental lexicon. They are operators on mental states" (Elman, 2011, p.16). Here, words are regarded as operators since their embeddings (or vector representations) directly alter the state of a system through weighted projection in connectionist frameworks. This type of connectionist views provides a better cognitive account for the context-dependent nature of human language processing given that a simple N-gram type model often fails to represent the entire context. It suggests that each word (bottom-up speech input) has activation values which can be mapped onto the system's current state defined by the activation pattern across the parallel processing units in the system's internal layer. Then, the state altered by the word is mapped onto the output units to constrain the upcoming continuations (see Rumelhart, Hinton & McClelland, 1987; Elman, 1990). Thus, the system's internal state changes as each word is heard incrementally

[1] If one context occurs more frequently in general than the other contexts, the potential candidates associated with the more frequent context expectedly co-occur more frequently than the candidates associated with less frequent context. For example, the potential candidates "treasure", "moment", "memory" and so on expectedly co-occur more with "love" than with "cherish" mostly because "love" occurs more frequently in general than "cherish"; not because "love" has more lexical association with those candidates than "cherish". Hence, calculating a conditional probability distribution (instead of a joint probability distribution) effectively adjusts for the difference in how frequently each context occurs in the natural language environment.

throughout a sentence. Until recently, training this type of model to learn the mappings between different layers in the architecture was not feasible due to technical limitations (e.g. processing speed and memory capacity of a computer). In this thesis, I directly test this connectionist account using a pre-trained neural network model based on 1 billion words (Jozefowicz et al., 2016) and compare its performance in explaining the variability in neural activity with computational models of linguistic constraints inferred from the human behavioural data.

1.3. Predictive computations involved in the multi-level speech processing: theoretical reviews

Understanding spoken language requires a complex set of perceptual and cognitive operations that transform the auditory input at the lexico-phonological level into a meaningful interpretation at the semantic and pragmatic levels. The field of psycholinguistics has long investigated the way that listeners perform such complex operations at these multiple linguistic levels. In light of the accumulating evidence, researchers have been arguing for and against the serial and parallel processing theories as described above. One of the theories supporting the serial processing view is so-called “syntax-first” theory. It is based on the notion that the human cognitive system is organized into a set of independent processing modules (Fodor, 1983), including a syntax module. The syntax module drives the initial interpretation of an upcoming word for syntactic structuring and a semantics module only plays a role during the later thematic assignment stage (Frazier, 1987). In contrast to this account, a parallel-interaction theory claims that it is the linguistic context and environment that guides the interpretation of each upcoming word through active interactions among multiple linguistic aspects including the syntax and semantics (Marslen-Wilson, 1975). Here, we briefly review these two conflicting views (the supporting evidence for these views were taken from both speech and reading domains).

Different views on “computation” in classical cognitive science

From the early 1960s, the field of cognitive science has made enormous efforts to understand the nature of the human mind and cognition. A dominant theory developed by Jerry Fodor proposed a view that cognition is a form of computation occurring in an information processing system called mind. One of the most important aspects of this theory, which explains its popularity in the early decades of cognitive science, is that a process of cognition can be mechanistically implemented with a given definition of computation. Perhaps, the most intuitive definition of computation is doing a mathematical calculation. So, what are the processes involved in mathematical calculation? It merely involves expressing a sequence of symbols in a way that does not violate the systematic relations among them; for example, $2 + 3 = 5$ involves five different symbols in a sequence in which each symbol can be manipulated under a set of mathematical rules (e.g. $3 + 2 = 5$). From this perspective, computation is merely a mathematical expression of a sequence of symbols.

However, there is no reason to restrict computation to a symbolic expression since we already know that we do numerous non-mathematical calculations in real-life engaging a variety of non-symbolic representational entities (a.k.a. representational vehicles; O'Brien, 1998). Hence, there is no reason not to generalize the concept of computation to a variety of cognitive processes such as planning, reasoning, attentional modulation and other executive functions outside a numerical calculation. Similarly, the systematic relations between the representational vehicles do not have to be defined only in terms of mathematical rules; for example, navigating to a particular location in a map not only requires identifying the symbols such as a black line representing a road, a red dot representing a traffic light and so on but it also requires combining them into a coherent picture of the objects in the represented domain (e.g. roads, traffic lights etc.). Consequently, the generalized definition of computation can be stated as "***a procedure in which representational vehicles are processed in a semantically coherent fashion***" (O'Brien, 1998).

So, how is "computation" mechanised? Or, in other words, how is a computational machine constructed? Most computationalists, including Jerry Fodor himself, defined computation in the light of Turing machines (a.k.a. Classicism). It involves a finite set of symbols encoding the information being manipulated in a semantically coherent fashion according to a finite number of rules. More specifically, the machine uses a tape with infinite memory capacity divided into a number of cells in which a symbol in a cell is modified and rewritten based on the table of rules. Then, it moves the position in the tape to either of the adjacent cells and continues the same process (the computation could be halted depending on the current position in the rule table). This renders the machine to behave in a way that complies to the computational instructions.

The core issue of implementing a Turing machine is to define the set of governing rules, or syntax. It requires systematic partitioning of the properties of a continuous variable (e.g. an input sentence). The semantically coherent partitions illuminate the syntactic structure which ultimately leads to the rule-governed behaviour of the machine in a discrete representational medium. In contrast to Turing machine, a newer branch of cognitive theory (known as connectionism) which emerged and gained popularity in the late 1980s introduced a densely interconnected network with vastly different architecture from the classical computational device. Here, the major distinction between them is centred on how semantically coherent behaviour is achieved: digital computation forces the behaviour to conform to syntax whereas analog computation relies on understanding a structural isomorphism (resemblance) between the representational vehicles and the objects in the represented domain (O'Brien, 1998).

While Turing machine was recognised with its simple architecture based on explicit statements of rules, the connectionist device was neurally inspired and recognised with the sophisticated interconnections between the processing units modulated by a set of weights (i.e. synaptic projection of action potentials between connected neurons). The modulatory weights are shaped through experience or learning, reflecting the neural development and cortical plasticity in humans. A variation in the representational vehicles, as a result, changes the processing state of the network via the weighted projection and such computation does not require any syntactic knowledge. It is rather claimed that the syntactic rules can be learned to a certain level only through statistical regularities among the vehicles which reflects natural acquisition of the first language in humans (Seidenberg, MacDonald & Saffran, 2002). See the section 4.2 in Chapter 4 and 2.3 in Chapter 2 for further explanations regarding the connectionist models.

The same line of debate emerged in the late 1970s in the field of psycholinguistics regarding the nature of computation during incremental speech comprehension in humans. Consistent with the classicist view that computation is guided by the syntactic rules, the syntax-first theory (Frazier & Fodor, 1978; Frazier, 1987) claimed that understanding speech starts with constructing the syntactic structure and guiding the interpretation of an upcoming linguistic unit under the minimal attachment principle. An opposing view known as constraint satisfaction (Mellish,1981; Altmann & Steedman, 1988) suggested that interpreting the upcoming unit is guided by the constraining source of information interactively at multiple linguistic levels.

1.3.1. Modular theory and "syntax-first"

Under a theory of human cognition in which the human mental architecture consists of autonomous modules only sensitive to domain-specific information (Fodor, 1983), Lyn Frazier (1987) developed the modular theory of language comprehension claiming that the autonomous modules of the language processor are key to understanding a sentence. The idea of "syntax-first" suggests that a syntax module initially constructs the simplest syntactic structure based on the grammatical category of each word, independent of its lexical-semantic information. Therefore, this theory is syntactocentric derivational, just like Chomskian theories of generative grammar described above. However, for syntax to guide the interpretation of an upcoming sentence, a straightforward paradox arises in relation to the basic intuition that speech comprehension is incremental (i.e. left-branching) because English grammar is right-branching (see Figure 1-1; Altmann & Steedman, 1988). To escape from this paradox, Frazier's theory was built upon the principles of minimal attachment and late closure (Frazier, 1987; Frazier & Fodor, 1978) which suggests that the input word is initially interpreted in a way that generates fewest phrase structure nodes (minimal attachment principle) and is reanalysed at a later stage if the actual sentence structure turns out to be different from the expected simplest structure. Hence, the presence of an autonomous syntax module enables listeners to interpret the input word in a particular way in the ambiguous settings (e.g. *"John told the girl that Bill liked the story"; "that" is typically interpreted as a sentential complementizer attached to the verb phrase instead of an adjectival clause attached to the complement noun phrase*) and to detect the syntactically preferred analysis even if it is pragmatically less plausible (e.g. *"a gift to a boy in the box"; interpreting the prepositional phrase "in the box" as being attached to "to a boy" instead of "a gift"*). Moreover, only if multiple possible interpretations have the same number of phrase structure nodes, the input word in a sentence is associated with the phrase being processed (late closure principle): for example, in "*The doctor said the patient will die yesterday*", the adverbial phrase "*yesterday*" tends to be attached to the verb phrase in the sentential complement "*die*", instead of the main verb phrase "*said*". In light of these principles, the syntax-first theory offers a practical solution to interpreting sentences with temporary structural ambiguities that constructs varying number of phrase structure nodes with a clear dominant and (a) subordinate(s) interpretations, known as "garden-path" sentences (Bever, 1970).

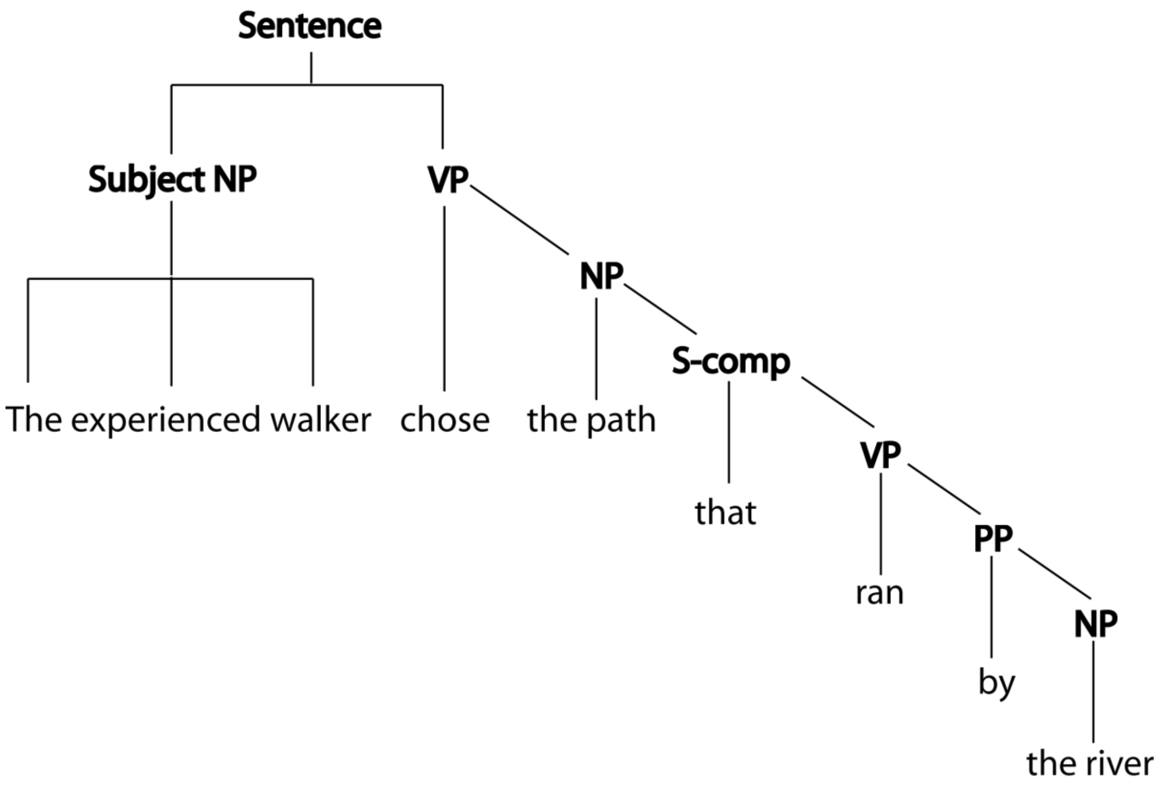


Figure 1-1: Syntactic parsing of an example sentence "The experienced walker chose the path that ran by the river", parsed by Link Parser online (http://www.link.cs.cmu.edu/link/submit-sentence-4.html). This simple figure illustrates that the branches of the syntactic tree expand towards right-side of the space. This is why English grammar is known as "right-branching" grammar. Abbreviations: NP = noun phrase, VP = verb phrase, S-comp = sentential complement and PP = prepositional phrase

To determine whether the syntax module autonomously processes a sentence at an early stage, a number of studies have tested whether the thematic role assignment of a preceding subject noun proceeds before the construction of its syntactic arguments. For example, Rayner Carlson and Frazier (1983) showed that pragmatic plausibility does not affect the initial syntactic analysis (Experiment 1):

a) The florist sent the flowers was very pleased
b) The performer sent the flowers was very pleased

Despite the fact that florists are expected to send flowers and performers are expected to receive them, the garden-path effect in their eye-movement data was observed for both sentences (based on the contrast with their unambiguous counterparts such as "*The performer* *who was sent the flowers was very pleased*"), suggesting that the pragmatic information did not affect the parser's initial choice of the simplest syntactic structure. Similarly, Ferreira and Clifton (1986; experiment 1) showed that the reading time at the underlined disambiguation marker of the following temporarily ambiguous sentences was not significantly different even though the thematic role of the subject noun (animacy) was manipulated:

c) The defendant examined **by** the lawyer turned out to be unreliable
d) The evidence examined **by** the lawyer turned out to be unreliable

The reading time was significantly higher in ambiguous sentences as above compared to unambiguous sentences like:

e) The defendant that was examined **by** the lawyer turned out to be unreliable

showing sensitivity to syntactic information but NOT to thematic role information in the early stage of sentence processing. In Experiments 2 and 3, they further demonstrated that reading times for Non-minimal attachment sentences (e.g. "The editor played the tape **agreed** the story was big") were longer than for Minimal attachment sentences (e.g. "The editor played the tape **and agreed** the story was big") only after the disambiguating word (underlined) regardless of context. These results were used as evidence to support the syntax-first theory.

However, these results were not replicated by later studies using the same manipulation (Spivey-Knowlton, Trueswell & Tanenhaus, 1993; Trueswell, Tanenhaus & Garnsey, 1994). In particular, Trueswell et al. (1994) showed that the processing difficulty that readers experience varies depending on the degree of semantic fit between an inanimate subject and a following verb; for example, it is difficult to process an upcoming "by" phrase that signals a reduced relative structure when the verb is thematically congruent with the subject noun as in c) compared to when it is not as in d). They suggested that the absence of such early semantic effects on interpreting the upcoming syntax is possibly due to methodological problems in Ferreira and Clifton (1986) such as weakly manipulated stimuli and uncontrolled difference in display mode between different conditions. Although the syntax first theory attracted attention in the field as a plausible explanation on some linguistic phenomena involved in interpreting syntactically complex sentences, the psycholinguistic evidence from behavioural studies has not been consistent.

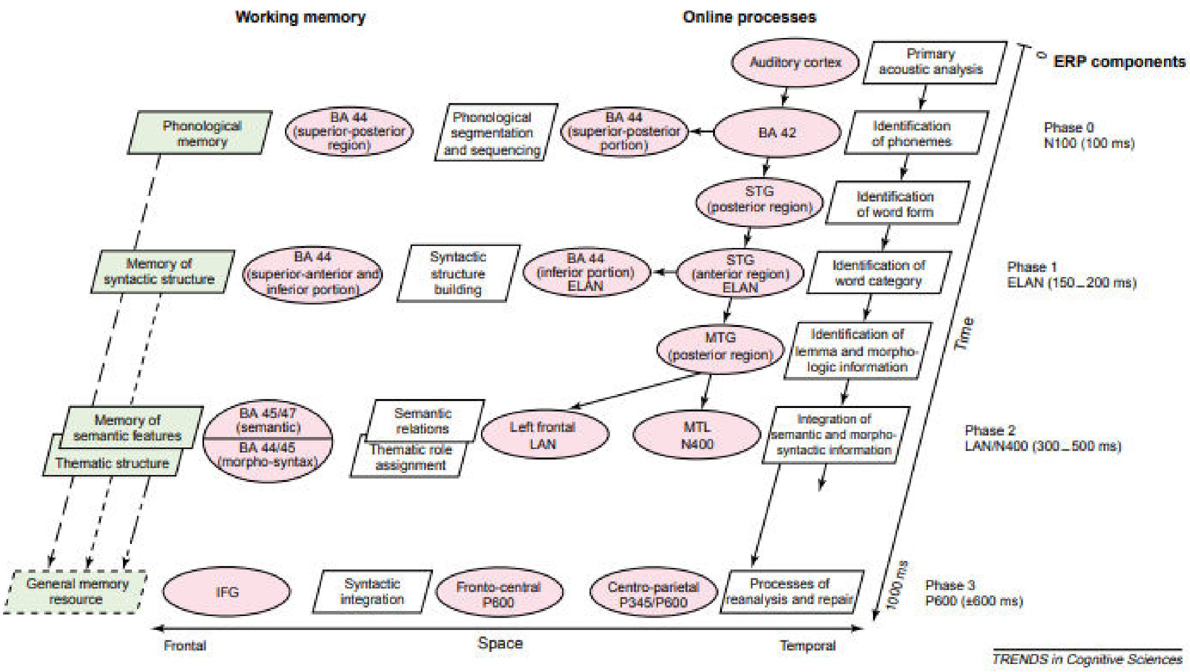
Evidence from ERP studies

Owing to the development of neuroimaging techniques, in the form of electroencephalography (EEG), which provides high temporal resolution , enabling researchers to capture the timing of syntactic and semantic processes during sentence comprehension, the syntax-first theory could start to be evaluated using time-sensitive information and thus get a clearer picture of the sequences of processing operations involved in language comprehension. The behavioural responses modelled in psycholinguistic studies such as error rates or reading times are static measures from which the cognitive processes associated with the experimental task (and stimuli) can be inferred. In contrast, EEG directly records the information processing in the brain as simultaneous real-time measures in millisecond resolution and an event-related potential (ERP) is a time-locked neural response to a stimulus summarized in the window in which mental operations occur.

By analyzing the evoked responses aligned to the onset of a word of interest (known as event-related potential (ERP) analysis), researchers found a number of ERP components (i.e. neurophysiological markers) associated with linguistic processing. First component, known as early left-anterior negativity (ELAN), occurs within 250ms after the onset of word-category violation (e.g. "*fed*" in "*The goose was in the fed*"; Friederici et al., 1993; Hahne & Friederici, 1999; Hahne & Jescheniak, 2001). This component was claimed to be very fast and highly automatic and was viewed as a neural index of the initial phrase-structure building based exclusively on the syntactic word-category information (Hahne & Friederici, 1999). This view is in line with the Fodorian modular theory of language processing and supports the autonomous role of the phrase-structure module at the initial processing stage of a word in a sentence. Another ERP component associated with morpho-syntactic errors such as verb-inflectional violation (e.g. "*mow*" in "*Every Monday, he mow the lawn*"), pronoun case violation (e.g. "*we*" in "*The plane took we to paradise and back*") and syntactic-gender violation in other languages, is known as left-anterior negativity (LAN), peaking around 400ms after the onset of violation (Gunter, Stowe & Mulder, 1997; Gunter, Friederici & Schriefers, 2000; Coulson, King & Kutas., 1998). In particular, Gunter et al. (2000) showed that the LAN was independent of their semantic variable, suggesting that the morpho-syntactic processing of a word is independent of its semantics at this stage.

In the ERP literature, many studies have investigated the integrative effect of context on the semantics of a target word. These studies have found a consistent and reliable ERP component associated with semantic processing which was originally reported by Kutas & Hillyard (1980). They found a negative ERP peaking around 400ms (so-called N400) after the onset of an improbable word with respect to the preceding context (e.g. '*dog*' in "*I take coffee with cream and dog*"). Later studies replicated and further generalized this N400 effect to natural language processing by showing that N400 amplitude is correlated with the degree of a word's expectancy based on the given context (Kutas & Hillyard, 1984; Kutas, 1993; Wicha, Moreno & Kutas., 2004; Delong et al., 2005; Federmeier et al., 2007; see also, Kutas & Federmeier, 2011; Frank et al., 2015).

The "Inter-modular interaction" was found at a later stage, reflected in a positive wave peaking around 600ms (P600) after the target word onset. This P600 component was originally reported as an index to syntactic anomaly, sensitive to the "Garden-path" sentences which require syntactic revision (e.g. "*The lawyer charged the defendant was lying*"; Osterhout & Holcomb, 1992; Osterhout, Holcomb & Swinney., 1994). Many studies have subsequently shown that this P600 component is less automatic and more controlled (Hahne & Friederici, 1999; Coulson et al., 1998) and possibly reflects interaction between syntax and semantics when syntactic re-analysis is required (Gunter et al., 2000). This is consistent with the Frazier's claim that inter-modular interaction depends on available computational resources, thus, requires the process to be more controlled. In the light of these studies, Friederici (2002) summarized the linguistic processing of a word as occurring in three phases, an initial phrase-structure building phase (ELAN), a subsequent morpho-syntactic (LAN) and semantic processing (N400) phase and a final revision phase (P600) at which different streams of information (i.e. syntax and semantics) are integrated (see Figure 1-2, taken from Friederici (2002)).


Working memory
Online processes
ERP components
0
Auditory cortex
Primary acoustic analysis
Phonological memory
BA 44 (superior-posterior region)
Phonological segmentation and sequencing
BA 44 (superior-posterior portion)
BA 42
Identification of phonemes
Phase 0
N100 (100 ms)
STG (posterior region)
Identification of word form
Memory of syntactic structure
BA 44 (superior-anterior and inferior portion)
Syntactic structure building
BA 44 (inferior portion) ELAN
STG (anterior region) ELAN
Identification of word category
Phase 1
ELAN (150–200 ms)
MTG (posterior region)
Identification of lemma and morpho-logic information
Time
Memory of semantic features
Thematic structure
BA 45/47 (semantic)
BA 44/45 (morpho-syntax)
Semantic relations
Thematic role assignment
Left frontal LAN
MTL N400
Integration of semantic and morpho-syntactic information
Phase 2
LAN/N400 (300–500 ms)
General memory resource
IFG
Syntactic integration
Fronto-central P600
Centro-parietal P345/P600
Processes of reanalysis and repair
1000 ms
Phase 3
P600 (±600 ms)
Frontal
Space
Temporal

TRENDS in Cognitive Sciences

Figure 1-2: Friederici's neurocognitive model of auditory sentence processing (2002). Taken from Friederici (2002).

A number of concerns have been raised about the early phrase-structure building component: ELAN (Steinhauer & Drury, 2012). First, the finding that this ELAN component was only observed in high visual-contrast condition (black font on white background) but not in low visual contrast condition (black font on grey background) (Gunter et al., 1999) and the finding that word-category violations with a particular preposition "*vom*" (*"by the"* in *"The white teeth were by the brushed"*) elicited N400 instead of ELAN (Gunter & Friederici, 1999) led to the problem of generalizing this component as a pure (modality-independent) syntactic component. Second, the timing of ELAN is expected to vary as a function of the input availability of word category information in speech. Assuming that the word-category information can only be accessed as soon as one recognizes the word, ELAN effects prior to a word's uniqueness point may reflect some other information or process. Finally, if there is a systematic difference between the two conditions in the ERP baseline prior to the target word onset, it often becomes difficult to fully attribute ELAN to the grammatical category violation of the target (e.g. contrasting "*Yesterday, I drank his brandy **by** the fire*" to "*Yesterday, I drank his **by** brandy the fire*" at the onset of "***by***" could already generate the confounding artefact at the onset between the conditions due to the difference in a preceding word).

In summary, the modular syntax-first theory provided an explanation about how humans process language: an autonomous syntax module guides the linguistic interpretation of a word in a sentence (e.g. in "*man bites dog*", the phrase-structure module constrains the way "*man*" and "*dog*" are integrated despite semantic implausibility).This theory attracted substantial attention in the field of neurolinguistics based on ERP evidence and provided a basis for Friederici's (2002) neurobiological model of language processing. Nevertheless, there have been a number of challenges to this view especially about its generalizability to natural language processing. Interestingly, most evidence comes from violation studies which rarely occur in a natural language environment. This is particularly problematic for interpreting neural activity as it could easily introduce non-linguistic, task-related confounds; for example, it could lead to the engagement of default mode network (DMN) which is activated (or deactivated) for task performance (Campbell & Tyler, 2018; Fox, Snyder, Vincent et al., 2005; Sormaz, Murphy, Wang et al., 2018).

The evidence from many of these ERP studies suffers from, at least, three major limitations. First, it often involves grammatical violations (especially the early syntactic components) which raise the possibility that these components may not be purely linguistic. Second, although the ERP components are consistently observed, a number of different interpretations have been made regarding the underlying cognitive operations of each component. In particular, interpreting the N400 has been controversial as there are many different factors that explain the variability in this component such as the frequency of a word's usage (Van Petten & Kutas, 1990), and the degree of a word's expectancy in a sentence (De Long, Urbach & Kutas, 2005). Some researchers have interpreted this component as an index of integration occurring after recognizing the target word (i.e. post-target process; see Brown & Hagoort, 1993) whereas the others have interpreted it as a reflection of facilitated activation of features associated with the target word (Lau, Phillips & Poeppel, 2008). Third, it is blind to the spatial dynamics in the brain. As described in section 1.4 below, there are regions and networks that are specifically involved in certain linguistic operations but the ERP analyses cannot elucidate the underlying generator of the components which makes the functional interpretations of each component even more difficult.

Illuminating the temporal dynamics of neurobiological processes involved in speech comprehension through empirical evidence has been one of the major research topics in the field of psycholinguistics. Previous ERP studies highlighted four different components (ELAN, LAN, N400 and P600) as neural markers of time-sensitive linguistic operations and discussed the underlying cognitive processes associated with each of these components (Figure 1-2). However, due to the experimental and methodological limitations described above, the validity and/or interpretability of these components in natural language environment has often been called into question. To address these issues, this thesis combine real-time neuroimaging (electro- and magneto-encephalography) with recent developments in multivariate statistics and computational linguistics to probe directly the dynamic patterns of time-sensitive neural activity that are elicited by spoken words, the constraints they generate on upcoming words, and the incremental processes that combine them into syntactically and semantically coherent utterance interpretations. In this way, I aim to directly test if information encoded in the spatiotemporal patterns of neural activity during natural spoken language comprehension is captured by the state-of-art computational models. The table below shows a summary of all models used in this thesis and their distinctive features that address a particular (set of) question(s).

The next section covers the contrasting theories of the syntax-first.

Table 1: summary of all models used in this thesis

	Modelled aspect of language	**Detailed descriptions**	**Notes**
Pretest-SCF (behavioural)	Syntactic constraint	See section 2.5.1 in Chapter2	*captures SCF preference of the entire context *participants' responses are categorized into 5 frames
VALEX-SCF (corpus-based)	Syntactic constraint	Korhonen, Krymolowski & Briscoe (2006)	* captures verbs' SCF preference *163 original SCF frames are collapsed into 5 frames
VALEX-WN (corpus-based)	Semantic constraint	See section 2.5.2(a) in Chapter2	*captures verbs' selectional preference in WN space *representation is optimized through MDL
LDA-DT (corpus-based)	Semantic constraint	See section 2.5.2(b) in Chapter2	*captures verbs' selectional preference based on their co-occurrence properties

LDA-WT (corpus-based & behavioural)	Semantic constraint	See section 2.5.2(b) in Chapter2 and section 3.3. in Chapter3	*captures selectional preference of the entire context *vector representation of each word from a pretest is blended by averaging
LSA (corpus-based)	Semantic content	Baroni & Lenci (2010)	*captures semantic representation of different nouns, verbs and adjectives based on their co-occurrence properties
LSTM0/1 (corpus-based)	See Figure 4-1 and 4-2 in Chapter 4	See section 2.3 in Chapter2 and section 4.2 in Chapter 4	*captures semantic properties of an input word at each incremental point in a sentence
LSTM-softmax (corpus-based)	See Figure 4-3 in Chapter 4	See section 2.3 in Chapter2 and section 4.2 in Chapter4	*captures semantic or syntactic constraints at each incremental point in a sentence

1.3.2. Parallel-interaction and constraint satisfaction approaches

Understanding speech engages many temporally overlapping processes at multiple linguistic levels. Therefore, it is critical to understand whether and when information at these multiple levels interacts so that the language system can efficiently constrain the rapidly unfolding words during speech comprehension. Unlike "syntax-first" models, active interaction among different linguistic levels to incrementally constrain an upcoming sentence is a central notion of parallel-interaction theories that have been shown in classical speech shadowing studies (Marslen-Wilson, 1973, 1975). In the light of this theory, Tyler and Marslen-Wilson (1977) further tested whether incrementally built semantic context actively guides the syntactic interpretation of syntactically ambiguous phrases:

f) If you walk too near the runway, landing planes …
g) If you've been trained as a pilot, landing planes …

In these examples, the phrase "*landing planes*" is syntactically ambiguous as it can be interpreted either as a gerundive phrase or as a noun phrase. If semantic context can guide the way that temporarily ambiguous syntactic phrases are interpreted, listeners will prefer to interpret “*landing planes”* in (f) as a noun phrase followed by a plural verb-form “*are*” whereas they will prefer to interpret it in (g) as a gerundive phrase followed by a singular verb-form “*is*”. By visually presenting a probe verb which was either congruent or incongruent to one or other semantically-constrained interpretation of the fragment and asking subjects to repeat the probe verb as quickly as possible, Tyler and Marslen-Wilson showed that the naming latencies for the incongruent probes were significantly longer than for the congruent probes.

Similarly, another experiment by Crain (1980) investigated whether the local garden-path effect in the following types of sentences can be controlled by referential context:

h) “The psychologist told the woman that he was having trouble with her husband”
i) “The psychologist told the woman that he was having trouble with to visit him again”

Both of these sentences are syntactically ambiguous at the point of “*that*” since each can be interpreted either as the opener to a complement clause or as the opener to a relative clause. In the experiment, these sentences were preceded by one of the following contexts:

j) “A psychologist was counselling a man and a woman. He was worried about one of them but not about the other.”
k) “A psychologist was counselling two women. He was worried about one of them but not about the other.”

(k) is a supporting context of (i), demanding relative-clause analysis because the relative clause “*he was having trouble with*” works as a modifier of a preceding noun phrase “*the woman*”, presupposing that there is more than one woman in the context and “*the woman*” in (i) refers specifically to one of them that the psychologist is having trouble with. On the other hand, (j) is a supporting context of (h) demanding complement-clause analysis due to the absence of such modifier in (i). Using a grammaticality judgment task with four conditions (2 x 2), Crain (1980) found an effect of context on processing syntactic ambiguities, consistent with the parallel-interaction theory. Moreover, this study provided evidence for the principle of referential success and failure (Crain & Steedman, 1985; Altmann, 1988), explaining that the complement-clause analysis in (i) can be discarded in the context of (k) because “*the woman*” as a simple noun phrase leads to referential failure (i.e. which one of the two women is “*the woman”* referring to?). Taken together, these studies demonstrated the influence of discourse (referential) context on the resolution of local syntactic ambiguities which cannot be explained by the syntax-first theory and minimal attachment principle.

Following on from these earlier studies (Marslen-Wilson, 1973, 1975; Tyler & Marslen-Wilson, 1977; Marslen-Wilson & Tyler, 1980), Mellish (1981) proposed a constraint-satisfaction account, claiming that readers/listeners incrementally evaluate referential relations among various objects (words or phrases) via continuously accumulating constraints which the referents (i.e. conceptual object such as place, entity, person etc.) of the referring expressions (i.e. words or phrases that refer to a particular referent) must satisfy. The set of "partially evaluated" referents becomes gradually refined as they progress through a sentence until a single candidate referent remains. Under this account, researchers investigated the way that comprehenders constrain the upcoming continuation.

MacDonald (1994) explicitly defined three different types of probabilistic constraint that might interactively resolve local syntactic ambiguity in the following sentences:

l) The patient heard the music (Active Transitive: Direct Object)
m) The patient heard with the help of a hearing aid (Intransitive: Prepositional)
n) The patient heard the nurses were leaving (Sentential complement)
o) The patient heard in the cafeteria was complaining (Reduced relative)

First, the thematic role of the early subject noun phrase "*the patient*" (i.e. it is more likely to be interpreted as a theme instead of an agent) already provides syntactic constraints (pre-ambiguity constraints). Second, a verb provides key information about its argument structure (verb subcategorisation constraints). Lastly, a direct object usually occurs immediately after the verb in English despite several exceptions. Hence, this "post-ambiguity" constraint inhibits an active transitive interpretation and helps specifically to resolve the "main verb/reduced relative" (MV/RR) ambiguity. The author demonstrated that all three constraints contribute to faster reading time and these constraints dynamically interact with each other such that ambiguity resolution was significantly facilitated when they converged compared to when they conflicted.

Consistent with this claim, a number of studies have shown the effects of context on interpreting the upcoming speech. For example, Tyler and Marslen-Wilson (1977) already demonstrated the context effects on syntactic interpretation of an upcoming sentence. Similarly, Spivey-Knowlton, Trueswell and Tanenhaus (1993) found an immediate effect of animacy of a subject (local semantic context) as well as the pragmatic and referential information from the discourse context on resolving the MV/RR ambiguity, supporting the interactive effect of context influencing syntactic interpretation (see also; Trueswell et al., 1994). Following on from these studies showing the early effects of discourse and local contexts on syntactic ambiguity resolution, it was further demonstrated that the lexical constraints of a verb strongly determines the syntactic interpretation of its complement. A verb naturally provides probabilistic information about a number of possible syntactic frames that can co-occur, known as subcategorization preference (Chomsky, 1964). A number of studies have firmly established that a verb's subcategorization preference directly influences the processing of its complement structure such that the reading time (or naming latency) is faster when the complement structure is preferred by the verb (Trueswell, Tanenhaus & Kello, 1993; Jennings, Randall & Tyler, 1997). Marslen-Wilson, Brown and Tyler (1988) also investigated the effect of verbal constraint on pragmatic, semantic and syntactic aspects of its argument and found that a verb exerts immediate influence on processing its argument in all of these aspects. In summary, all of these studies demonstrate the importance of both lexical and contextual constraints and their interaction (e.g. when they are in conflict vs. when they converge) in order to guide the syntactic interpretation of an upcoming sentence.

Such interaction between contextual and verbal constraints in constraining referential pronouns has also been reported by Marslen-Wilson, Tyler & Koster (1993). They explicitly manipulated the referential context and the main verb of a subject pronoun of a target sentence as following:

p) After the surgeon had examined the 12-year-old girl with the badly broken leg, he decided he would have to take immediate action. He'd had a lot of experience with serious injuries. He knew what he had to do next.
    a. He quickly injected … [probes: him or her]
    b. She quickly injected … [probes: him or her]
    c. Quickly injecting … [probes: him or her]

In this stimulus, both context and verb agree that "surgeon" is the agent, preferring 'He' as a subject pronoun in p)-a. Now, consider another stimulus:

q) Mary lost hope of winning the race to the ocean when she heard Andrew's footsteps approaching her from behind. She was slowed down by the deep sand. She had trouble keeping her balance.

a. She overtook … [probes: him or her]
b. He overtook … [probes: him or her]
c. Overtaking … [probes: him or her]

In this stimulus, the agent in the context (Mary) is in conflict with the agent preferred by the verb “overtook” (Andrew). Their results showed that different kinds of processing information are flexibly adapted to link utterances to discourses. In the agreement condition as in p), they found that the naming latency for the probe “*her*” was fastest in p)-a than in any other target sentences with different probes including p)-c with “*her*” as a probe, demonstrating the importance of both contextual and verbal constraints. In the conflict condition as in q), there was no difference in naming latency between different probes for q)-a but “*her*” was named significantly faster in both q)-b and q)-c, emphasising the verbal constraint as a primary source of constraining the thematic role of its noun phrases.

These psycholinguistic studies find evidence for the parallel-interactive and incremental nature of the human language processing system which is key to resolving both syntactic and semantic ambiguity. In line with these studies, Altmann and Steedman (1988) proposed that the human language processor is “parallel fine-grained weakly interactive” such that various kinds of linguistic constraints are represented in parallel and are incrementally adapted and refined as the processor progresses through a sentence. Besides, the term “weakly-interactive” implies that the referential context interactively disposes certain analyses that are proposed by syntactic knowledge; for example, two different syntactic interpretations of “landing planes” are proposed by syntactic knowledge (i.e. gerundive vs. noun-modifier phrase) and the preceding context interactively disposes certain interpretations that are semantically incompatible (see f) and g) as examples). This is in line with the view of prediction as a graded and probabilistic phenomenon (Kuperberg & Jaeger, 2016) such that “weak-interaction” may lead to a shift towards a particular dimension in the probability distribution but never really “rule-out” other dimensions in a serial manner.

Neuroimaging evidence

To address issues of incrementality in language comprehension and investigate the neural underpinnings of the kinds of phenomena described above, we need to consider data from studies using MEG. An important advantage of MEG neuroimaging data (compared to behavioural data) is that it enables researchers to investigate the spatial and well as the temporal dynamics of various constraints and potential interactions among them as well as the effects of such constraints on the processing of a subsequent word. In so doing, it can overcome some of the limitations of EEG which has typically not provided spatial information about language processes. Despite abundant evidence from many studies in the ERP literature for such effects of contextual and verbal constraints (Hagoort, Hald, Bastiaansen & Petersson., 2004; Nieuwland & Van Verkum, 2006; Bicknell et al., 2010), ERP analysis is specific to a predefined time-window, time-locked to the onset of an event. Since the representational content of neural activity varies as a function of time during speech comprehension, an advance is to analyse the time-course of source-localized brain recordings and to model how cognitive information changes over space and time from constraint to integration. This is particularly essential for modelling the constraints, often in a multidimensional space (i.e. simultaneous representations of potential candidates).

To my knowledge, there are only a few such studies in the literature (Tyler et al., 2013; Kocagoncu et al., 2017; Klimovich-Gray, Tyler, Randall, Kocagoncu, Devereux & Marslen-Wilson, 2019) and none of them looked into how the brain processes syntax in the presence of the discourse context. Importantly, Tyler et al. (2013) showed that the strength of a verb's subcategorisation preference for a direct object frame is crucial in interpreting syntactically ambiguous phrases like *"... juggling knives ..."*

r) "*In the circus, juggling knives is less dangerous than eating fire*" (preferred)
s) "*In the circus, juggling knives are less sharp than people think*" (unpreferred)
t) "*There are many reasons why boiling liquids are to be handled carefully*" (preferred)
u) "*There are many reasons why boiling liquids is an effective way to kill germs*" (unpreferred)

The disambiguation word "*is*" or "*are*" clarifies whether the ambiguous phrase is a noun phrase or a gerundive phrase and initiates reanalysis if it turns out to be inconsistent with listeners' expectations. Their results showed that the degree of preference varied as a function of the direct object preference of the verb in the ambiguous phrase, consistent with constraint satisfaction but not with minimal attachment. They found that such direct object preference information is encoded in left middle temporal gyrus as soon as the verb is pronounced lasting about 110ms whereas left inferior frontal gyrus (LIFG) was sensitive to reanalysis from 374ms to 714ms, demonstrating the differential roles of frontal (reanalysing) and temporal (constraining) regions in resolving syntactic ambiguities. These results are consistent with the constraint satisfaction account. The detailed analysis pipeline and its appealing characteristics (as well as limitations) are described in Chapter 3. However, further research is needed to corroborate the underlying neural mechanism of interactive predictive processing of incremental speech.

1.4. Neuroanatomy of speech processing

Studying behaviour alone gives a limited picture of the kinds of cognitive operations which underlie the comprehension of spoken language. Rather than inferring such cognitive operations from the output (behaviour) of a system (brain), we can look directly inside the system and find out the regions and networks involved in cognitive functions. Hence, this section is designed to address the question about "where" in the brain the various linguistic processes at different levels take place based on experimental evidence from neuropsychological patients, positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) studies with healthy people. In particular, I focus on syntactic and semantic processing in the brain, the two central cognitive operations for understanding a message from a linguistic input.

Studying brain-damaged patients is especially important for establishing a causal link between brain and behaviour. Early classical neuropsychological studies found that brain-damaged patients having difficulty producing grammatical sentences commonly had damage to Broca's area (left posterior inferior frontal gyrus corresponding to BA44 and 45) whereas those having difficulty understanding meaning had damaged Wernicke's area (left posterior superior temporal gyrus corresponding to BA22). In light of these early patient studies, the classical neurobiological model known as Wernicke-Lichtheim-Geschwind (WLG) model (Geschwind, 1965) proposed that the human language faculty is situated in the left perisylvian cortex with a strict division of labour between Broca's and Wernicke's areas which are anatomically connected via the arcuate fasiciculus (see Figure 1-3). In this way, the functional role of each region in cognitive processing can be studied. However, the crucial limitation of patients study is that, although all patients may commonly have a lesion in a particular brain region, the extent of the lesion varies across different patients, which may lead to other cortical dysfunctions and behavioural deficits. Due to this reason, it is often very difficult to obtain proper samples to study that have maximally consistent and focal lesions in the brain region of interest.

More recently, owing to the development of neuro-imaging techniques, especially fMRI, non-invasive tracking of localized brain activity has become viable which has allowed researchers to investigate the brain regions involved in different levels of linguistic processing. Using this method, the neurobiological system in which the various linguistic computations are instantiated has thoroughly and extensively been investigated in healthy subjects. In this section, I review recent models of the neurobiology of language, providing deeper insights into how the brain processes different aspects of linguistic information from the neuro-imaging studies of both patients and healthy subjects.

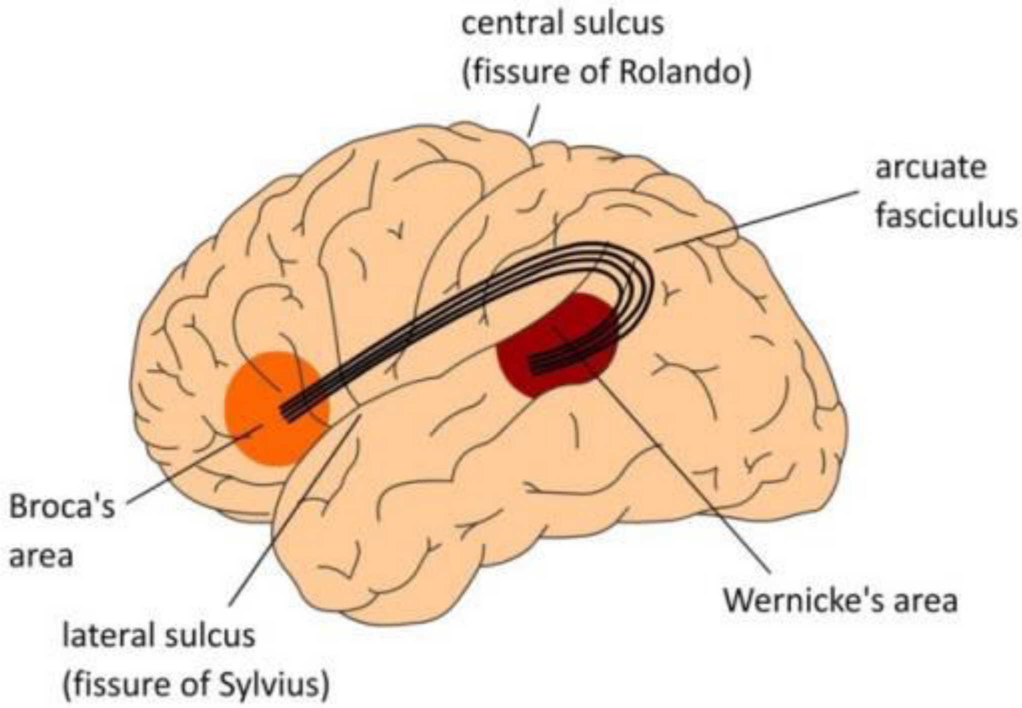


Figure 1-3: Visualisation of Wernicke-Lichtheim-Geschwind (WLG) model taken from Hagoort (2013).

1.4.1. Syntactic and semantic processing in the brain

Understanding a sentence requires interpreting each word in a syntactically and semantically coherent way. Many studies have investigated the process of combining words into a syntactically coherent sentence during natural language comprehension by manipulating the degree of complexity in syntactic combinations. For example, the degree of syntactic complexity and local ambiguity varies across different sentences in a natural language environment. Several studies have manipulated the syntactic complexity of sentences without violating the grammar (Just, Carpenter, Keller, Eddy & Thulborn., 1996; Caplan, Alpert & Waters., 1998; Caplan, 1999) to investigate the neural substrates of syntactic processing in healthy subjects. Consistent with the traditional view of Broca's area as a "syntax-region", these studies showed that syntactically more complex sentences (e.g. object relatives structure like "*The reporter who the senator attacked admitted the error*") elicited stronger activation in Broca's area (L-BA44/45) than less complex sentences (e.g. subject relatives structure like "*The reporter who attacked the senator admitted the error*"). Moreover, Rodd et al. (2010) manipulated the syntactic ambiguity of a sentence similar to that in Tyler & Marslen-Wilson (1977) and Tyler et al. (2013):

a) He noticed that landing planes frightens some new pilots (high-ambiguity)
b) She thought that renting flats requires a large deposit (low ambiguity)

In an fMRI study of healthy subjects, they found that the posterior portion of LIFG and LpMTG were strongly activated for high-ambiguity sentences compared to low-ambiguity sentences. In conjunction with the evidence from the morphological studies, these studies emphasized the functional role of the commonly activated left-lateralized LIFG-LpMTG network in syntactic processing at both lexical and sentential levels.

In contrast to the regions involved in syntactic processing, a more bilateral and distributed network is involved in semantic processing including temporal cortex, inferior parietal cortex and inferior frontal cortex (Binder et al., 2009; Price, 2010, 2012). To investigate brain regions involved in semantic processing, a number of fMRI studies have contrasted the brain activity associated with semantically plausible and implausible sentences. For example, Roder, Neville, Bien & Rosler (2002) found that meaningful sentences elicited greater activation in perisylvian cortex including LIFG and both anterior and posterior temporal regions than pseudo-word sentences (stronger in left). This pattern of results has been observed in other studies using similar experimental manipulations (Narain, Scott, Wise, Rosen & Leff., 2003; Crinion et al., 2003).

The functional role of posterior temporal regions in speech comprehension has long been demonstrated by patients studies (Bates, Wilson, Saygin et al., 2003; Gorno-Tempini, Dronkers, Rankin et al., 2004) and by neuroimaging studies on healthy subjects (Binder, Frost, Hammeke et al., 1997; Miglioretti & Boatman, 2003). Moreover, several studies have reported the engagement of posterior ITG in spoken word processing (Binder et al., 2000) and semantic ambiguity resolution (Rodd et al., 2005). For example, by manipulating the semantic ambiguity of a spoken word in a sentence, Rodd and colleagues (2005) found increased activation in bilateral anterior IFG (BA45) and left posterior ITG when processing semantically ambiguous sentences (e.g. *"She saw a hare/hair while she was skipping across the field*) compared to unambiguous sentences. Taken together, these studies suggest that the posterior temporal lobe is involved in the lexical-semantic processing of a spoken word with or without context. In conjunction with the abundant evidence for the involvement of Heschl's gyrus (HG) and posterior STG in acoustic-phonetic processing (Naatanen, Lehotokoski, Lennes et al., 1997; Morosan, Rademacher, Schleicher et al., 2001; Formisano, Kim, Di Salle et al., 2003; Mesgarani, David, Fritz & Shamma, 2008), these studies showed evidence for a functional role of this posterior temporal region as a phonological-semantic interface (see Hickok & Poeppel, 2004; 2007).

Also, another consistently reported region in semantic processing studies is the bilateral inferior frontal gyrus. For example, Kang, Constable, Gore and Avrutin (1999) investigated the brain regions involved in processing two-word phrases in one of the three conditions (normal, syntactically anomalous or semantically anomalous) in an fMRI study without an explicit task. They reported significant activation in bilateral IFG when processing semantically anomalous phrases whereas syntactically anomalous phrases elicited activation only in LIFG (L-BA44). More recent studies have reported that the strong activity in bilateral IFG reflects increased semantic competition or conflicting semantic information inconsistent with the semantic constraints (Vartanian & Goel, 2005; Peelle, Troiani & Grossman, 2009). For example, a spoken word recognition fMRI study which varied the degree of cohort competition (a number of competing word candidates) showed significant activation of bilateral anterior IFG (BA45/47) with increased cohort competition (Zhuang, Tyler, Randall, Stamatkis & Marslen-Wilson, 2012). Given that increased activation in this region has also been observed for processing semantically ambiguous sentences (Rodd et al., 2005), this bilateral IFG region may play an important role in semantically constraining the target word based on the context, selecting the likely candidates and integrating the target into the context.

According to Binder et al. (2009), the most consistently reported region across 120 functional imaging studies regarding semantic processing is the left angular gyrus (LAG) located in the inferior parietal cortex. For example, Obleser and Kotz (2009) showed that

LAG activation was only observed when successful speech comprehension was accomplished either by increased signal quality or by strong semantic constraints. Activation in LAG was reported when semantically anomalous words were embedded in a sentence (Ni, Constable, Mencl et al., 2000) and when processing a coherent narrative compared disconnected sentences (Xu, Kemeny, Park, Frattali & Braun, 2005). Similarly, Humphries, Binder, Medler and Liebenthal (2007) showed that processing semantically coherent sentences elicited activity in LAG compared to semantically random sentences (e.g. "*The freeway on a pie watched a house and a window*"). A recent single word recognition study demonstrated that trial-wise variability both in the degree of cohort competition and in the ease with which the semantic features are integrated generated patterns consistent with multivariate activity patterns in LAG (Kocagoncu et al., 2017). These various lines of evidence suggest that this region is involved in conceptual representation and integration at word, sentence and discourse levels.

Lastly, the anterior temporal lobe (ATL) has been suggested as a core semantic processing region from studies of patients with semantic dementia (Mummery, Patterson, Price et al., 2000; Gorno-Tempini, Rankin, Woolley et al., 2004), a virtual lesion study using repetitive transcranial magnetic stimulation (picture naming and word comprehension; Pobric, Jefferies & Ralph, 2007), a meta-analysis of 97 functional imaging studies elucidating a functionally unified bilateral ATL system (Rice, Lambon-Ralph & Hoffman, 2015). Other functional imaging studies which contrasted the neural activity between sentences and word-lists (or sounds) also showed strong activation in bilateral anterior (superior/middle) temporal regions (Mazoyer, Tzourio, Frak et al., 1993; Schlosser, Hutchinson, Joseffer et al., 1998). Consistent with these results, ATL is involved in syntactic structure building during natural language comprehension (Brennan, Nir, Hasson et al., 2012) and damage in this region has been associated with deficits in understanding complex syntactic structures (Dronkers, Wilkins, Van Valin et al., 2004), suggesting the role of this region in combinatorial processing in natural language comprehension (Hickok & Poeppel, 2007). Rogalsky and Hickok (2008) tested if activity in ATL is modulated by syntax, compositional semantics or both using a selective attention paradigm with an error detection task (either syntactic or semantic). By specifying the sentence-specific ATL region responding to sentences compared to noun-lists, they showed that this region is sensitive to both syntactic and compositional semantic functions (except for a small proportion of this area that is only sensitive to semantic functions).

In summary, semantic processing during language comprehension recruits an extensive bilateral fronto-temporo-parietal network in contrast to syntactic processing which involves a left-lateralized fronto-temporal network. Inside this extensive network, four different regions including bilateral posterior and anterior temporal, inferior frontal and left inferior parietal areas (LAG) have consistently been reported. From these studies, the functional role of each of these areas has been suggested; 1) the posterior temporal regions are involved in lexical analysis of a word by mapping it onto its semantics, 2) the inferior frontal regions are involved in resolving competitions during the process of constraining the interpretation, 3) LAG is involved in representation of conceptual semantics and 4) the anterior temporal regions are involved in combinatorial processing such as semantic composition. In the following section, I describe a number of neurobiological models of language processing in humans which are built upon the rich evidence from these studies.

1.4.2. Anatomical connectivity within language networks

As suggested in the classical Wernicke-Lichtheim-Geschwind model, the neuroanatomical connectivity between frontal (Broca's area) and temporal (Wernicke's area) regions (arcuate fasciculus in figure 1-3) is crucial for preserving the flow of information between these regions and damage in this white matter tract is known to result in conduction aphasia characterized by repetition difficulty (Tanabe, Sawada, Inoue et al., 1987). The recent development of diffusion imaging techniques has revealed much richer white matter connectivity between these areas, organized into a dorsal route (superior longitudinal fasciculus (SLF) and arcuate fasciculus (AF)) and a ventral route (extreme capsule (EC) and uncinate fasciculus (UF)). The well-known technique for mapping white matter tractography in the brain is called diffusion tensor imaging (DTI) which measures the diffusivity of water molecules influenced by the microscopic architecture of the brain tissue (orientation of myelinated axon fibres). Using this approach, Catani and Jones (2005) calculated the diffusion index at each voxel known as fractional anisotropy based on the eigenvalues of a diffusion tensor matrix (3 x 3 matrix of diffusion anisotropy) and produced a brain map of fractional anisotropy. Their results first revealed two distinctive dorsal routes: direct white matter connectivity between Broca's and Wernicke's areas and indirect white matter connectivity between these regions via the inferior parietal lobe. Similarly, three distinct pathways between these regions have been reported including SLF, EC and UF (Anwander et al., 2006).

The functional significance of these pathways in syntactic processing has been controversial. Friederici, Bahlmann, Heim, Schubotz & Anwander (2006) suggested that the dorsal and the ventral pathways are functionally segregated based on results from an artificial grammar study. They claimed that the ventral route connecting the frontal operculum (FOP) to anterior STG via UF is involved in analysing transitional structures (e.g. ABAB sequence) whereas the dorsal route connecting Broca's area to posterior STG/STS supports the analysis of hierarchical structures (e.g. A[AB]B sequence). Based on the evidence that aLIFG (L-BA47/45) is involved in semantic processing (Gough, Nobre & Devlin., 2005; Vigneau, Beaucousin, Herve et al., 2006), Friederici (2009) suggest that another ventral route, EC, supports semantic processing.

In contrast to this claim, Rolheiser, Stamatakis & Tyler (2011) scanned patients with left-hemispheric lesions and correlated the fractional anisotropy voxel-by-voxel with their language test scores for the comprehension and production of phonology, morphology, syntax and semantics. Their results revealed that the comprehension test scores for phonology and morphology were correlated with the dorsal pathway (phonology: aAF adjacent to the precentral gyrus and supramarginal gyrus; morphology: AF/SLF near BA39/40), those for semantics were correlated with the ventral pathway (pEC near pMTG) and those for syntax were correlated with both pathways (pAF near supramarginal gyrus and tracts near LIFG and temporal pole). From these results, they emphasised the synergistic role between the ventral and the dorsal streams for linguistic processing, depending on the "varying demands of different components of language function". Another study (Griffiths, Marslen-Wilson, Stamatakis & Tyler., 2013) carried out probabilistic tractography analyses on patients and controls using LIFG and LpMTG as seed clusters based on a previous patient study (Tyler, Wright, Randall, Marslen-Wilson & Stamatakis., 2010). Their results corroborated the causal role of both dorsal (AF) and ventral (EC) pathways in syntactic processing.

The results regarding the functional role of each white matter tract must be indirectly interpreted with respect to the role of grey matter regions connected by the tract because the white matter tract itself does not produce behaviour. In other words, ascribing functions to white matter tracts require rich understanding of the grey matter regions that it connects. As a result, any dysfunction or syndrome associated with lesions specifically in a white matter tract is attributed to disabled communication between the regions; for example, conduction aphasia, consistently observed with lesions in arcuate fasciculus, is caused by disabled communication between Broca's and Wernicke's areas (Hickok & Poeppel, 2004). However, understanding various functions enabled by the anatomy of white matter tracts is very complex as Catani and Jones (2005) already showed that there can be multiple pathways in a tract; the dorsal AF tract involves an indirect pathway that passes through temporo-parietal junction (TPJ) to connect between Broca's and Wernicke's areas. These studies emphasize the importance of understanding the interactive nature of communication within an extensive language network. Therefore, further studies should evaluate the changes in the functional connectivity over time between different regions against the white matter anatomy while processing a linguistic input. This will illuminate the interactive nature of neural communication among different regions in the language network for various incremental computations associated with analyzing, constraining and integrating each word during incremental speech comprehension.

1.4.3. Neurobiological models of speech comprehension

The results from above studies described in 1.4.1. and 1.4.2. have engendered a broad agreement that speech comprehension requires a widely distributed bilateral fronto-temporo-parietal network. However, the exact functional roles of sub-regions and networks within this broad language network are controversial (e.g. locus of syntactic processing). In this section, I briefly review the general consensus and conflicts between different neurobiological accounts of syntactic and semantic processing during speech comprehension.

In the neurolinguistic literature, the dual-stream model of speech comprehension gained much attention (Hickok & Poeppel, 2000, 2004, 2007). This model proposes two functionally segregated streams in the brain when processing speech input. First, the ventral stream maps the acoustic-phonetic information of the speech input onto the conceptual and semantic representation via bilateral pMTG/pITG. After lexical-semantic analysis in these regions, the combinatorial processing of the speech input sequentially takes place in ATL as described in Rogalsky and Hickok (2008). Second, the dorsal stream which also takes the acoustic-phonetic input projects to temporo-parietal junction at the sylvian fissure (called area Spt) and enables auditory-motor integration (see also, Saur et al., 2008). The anatomical connectivity to pLIFG via the dorsal route (SLF/AF) supports articulatory processing.

Conflicting with the view above, other accounts have suggested pLIFG as a locus of syntactic processing (Friederici, 2009, 2011, 2012). More specifically, Friederici (2009) suggested a part of the ventral route connecting anterior STG to FOP via UF is involved in local phrase structure building whereas a part of the dorsal route connecting L-BA44 (pars opercularis) to posterior STG via AF/SLF engages complex hierarchical syntax (e.g. long distance dependencies). She has also claimed that posterior STG is a locus of both syntactic and semantic integration as it receives input from L-BA44 via the dorsal route (syntax) and a number of semantic regions including posterior MTG, LAG and L-BA45/47 (Grodzinsky & Friederici, 2006; Friederici, 2011, 2012). In contrast to Friederici's claim (2009, 2011, 2012) that a part of the ventral pathway (EC) is involved only in semantic processing, some other views have emphasised the synergistic interaction between LIFG (BA44/45) and LpMTG (Tyler & Marslen-Wilson, 2008) via the dorsal (AF/SLF) and the ventral (EC) pathways (Rolheiser et al., 2011; Griffiths et al., 2013).

Consistent with this view, Hagoort's "Memory, Unification and Control (MUC)" model (2005, 2013) suggests that LpMTG is involved in retrieving lexico-syntactic information from "Memory" (lexicon) describing the local syntactic preferences of a lexical item (e.g. verb's subcategorisation information). This model also suggests that LIFG (BA-44/45) is involved in "Unification" which refers to the process of combining elements to derive new and complex meaning (i.e. integrated representations). On top of the evidence that the left posterior temporal cortex is involved in lexical processing (see Hickok & Poeppel, 2004), direct evidence was given by Tyler et al. (2013) who showed that a verb's local syntactic preference for a specific frame (i.e. direct object) is represented in LpMTG from the offset of the verb. Moreover, they also showed that the activation pattern of LIFG (BA45) across different sentences varies as a function of the presence of syntactic ambiguity and sensitivity to reanalysis due to being garden-pathed. Hagoort (2013) emphasised the dynamic interplay between "Memory" (LpMTG) and "Unification" (LIFG) such that LIFG unifies syntactic information retrieved from LpMTG for selective pre-activation (Snijders, Vosse, Kempen et al., 2008; Snijders, Petersson & Hagoort 2010). Furthermore, he argued for a functional subdivision in LIFG into an anterior portion (L-BA45/47) involved in semantic unification and a posterior portion (L-BA44/45) involved in syntactic unification. This claim conflicts with the accounts mentioned above that domain general combinatorial processing takes place in ATL (Hickok & Poeppel, 2007) or LpSTG (Friederici, 2011, 2012).

Bornkessel-Schlesewsky and Schlesewsky (2013) proposed a different view which resolves these conflicts to a certain extent: the ventral stream (ATL) engages the time-independent computation, namely the unification of conceptual schemata (incorporating one schema into the slot of another) whereas the dorsal stream (LIFG-LpSTG/MTG) is involved in time-dependent processes such as prosodic segmentation, syntactic structuring and understanding internal thematic relations. Therefore, the time-independent conceptual schemata allow listeners to track and develop the sentence-level (or even discourse-level) representation by closely interacting with the time-dependent processes of identifying and cumulating the incrementally unfolding words. Correct identification of schemata requires understanding the sentence structure and unification of schemata occurs at the phrasal and sentential levels, explaining why ATL is involved in multi-aspect combinatorics. They also suggested that LIFG is involved in conflict resolution and general cognitive control which projects back to the posterior temporal regions via the dorsal route (AF/SLF) for syntactic structuring and identifying internal thematic relations (see also, Bornkessel, Zysset, Friederici, Von Cramon & Schlesewsky., 2005; who showed that activity in posterior STS reflects the complexity of verb-based argument hierarchy whereas LIFG (BA44/45) activity reflects linearization demands on processing hierarchical structures).

In summary, these neurobiological models illuminate the potential explanations of how the brain processes different aspects of language during incremental sentence comprehension. However, the majority of these accounts do not provide predictions about the temporal dynamics of brain activity associated with these different aspects of computations at word, phrase and sentence levels. Given the incremental nature of speech comprehension, it is critical for neurobiological models to explain the temporal dynamics of incremental computations in the brain as well as its spatial dynamics. In particular, not many studies have investigated the predictive nature of incremental computations in the context of a natural sentence comprehension. Although many previous studies proposed that the bilateral inferior frontal and anterior temporal areas are involved in constraining and combining the target word, none of them based on the fMRI or PET results could capture how the computational properties that these regions represent change over time as each word is incrementally unfolding in a sentence. Here, this dissertation aims to enrich the understanding of the temporal progression of constraints and integration (syntactic and semantic) both at word and at context (sentence) level.

1.4.4. The lexicalist approach

Most current linguistic theories assume that the lexical properties of a verb or other predicate that heads the sentence strongly determines the syntactic interpretation of the overall structure of the argument phrase (e.g. in a sentence “The child tried to find the picture”, the infinitival frame is strongly activated by the verb “tried”). This is the main assumption of the lexicalist accounts which is manifested in many grammar theories. Especially, the strong version of this account claims that the grammatical and semantic information localized within lexical entries is used to constrain the upcoming linguistic unit (Sag & Wasow, 2011). Hence, this account suggests that constructing the sentence-level representation is associated with every lexical item in the sentence because simple grammatical structures are easily derivable from lexical constraints as in SCF which is also endorsed by theta role assignment in lexical functional grammar (Bresnan, 2001).

Consistent with a strong parallel-interaction account of language comprehension (see above), this lexicalist account claims that understanding a sentence requires activating the lexical properties of incrementally unfolding words and constraining the way that upcoming predicate arguments are interpreted by close interactions among different levels of processing dimensions. The fact that lexical properties in the context can guide the syntactic interpretation interactively from the early stage of processing is in contrast to the syntax-first theory which emphasizes the use of explicit syntactic knowledge independent of the lexical-semantics at the initial processing stage (Frazier, 1987). Consistent with the lexicalist claim, previous psycholinguistic studies showed significant influence of verb’s lexical constraint on processing the upcoming words (Marslen-Wilson et al., 1988; Trueswell et al., 1993; Jennings et al., 1997; Hare, McRae & Elman, 2003).

Many neurobiological models of speech comprehension agree that posterior temporal regions are involved in activating lexical information (Hickok & Poeppel, 2007; Hagoort, 2013). For example, Tyler and colleagues (2013) showed that the SCF information associated with a preceding verb is activated in LpMTG as soon as the offset of the verb lasting for about 110ms in their source-localized MEG study. This is consistent with the claim of MUC model (Hagoort, 2013) that LpMTG is involved in “memory” function of activating this lexico-syntactic information. Moreover, the bilateral posterior STG/MTG regions extending to ITG were suggested to form a “ventral stream” in which the phonemically identified speech input is mapped onto its lexico-semantic representation (Hickok & Poeppel, 2004, 2007).

Consistently, these areas have been commonly reported in other neuroimaging studies of lexical and semantic processing as reported in reviews and meta-analysis (Binder et al., 2009; Price et al., 2010). Therefore, if the lexicalist claim explains human speech comprehension, the posterior temporal areas are likely to be activated for representations of lexical properties such as verb-based SCF soon after the lexical item is recognized until the onset of a word that reveals the actual frame.

1.5. Issues addressed in this thesis

However, a question still remains; to what extent can lexically-based constraints explain the predictive processing in the brain during natural speech comprehension? For example, other studies showed how the semantics/pragmatics of the entire preceding context can influence the syntactic and semantic interpretations of a phrase (see Tyler & Marslen-Wilson, 1977; Marslen-Wilson et al., 1993). As described above, the important notion of a constraint-satisfaction theory is the accumulative nature of constraints from which the upcoming input is evaluated. Unlike the lexically-derived constraints, the bilateral ATL is likely to be involved in representation of accumulative constraints according to the neurobiological models, as these regions operate combinatorial processing (Hickok & Poeppel, 2007; Rogalsky & Hickok, 2008; Bornkessel-Schlesewsky & Schlesewsky, 2013). Nevertheless, the incremental changes in the dynamic representation of such cumulative constraints in the brain have not been thoroughly investigated. In this thesis, I address this issue by developing a number of constraint models based on different theoretical assumptions in order to address the central issues outlined below:

First, the following three questions are addressed in Chapter 3 using syntactic and semantic models of constraints and integration, either based on a verb or a full preceding context. It is an extensive chapter in which a number of different models were tested and evaluated to address these questions:

1) *What are the linguistic bases of predictive computations?*

Following on from this discussion, I investigate how well models of constraints based on the full preceding context (contextual constraint) or based solely on the preceding verb (lexical constraint) perform in explaining the variability in spatiotemporal dynamics of neural activity.

2) *Are syntactic constraints activated prior to the activation of semantic properties in order to enable early phrase structure building before constraining the lexical-semantics?*

Activating all relevant information in parallel in order to constrain the upcoming speech is the main idea of the parallel-interaction theory (Marslen-Wilson, 1975). I evaluate this model by testing the syntactic and semantic models of constraints, and compare the earliness with which they are activated in the brain.

3) *Do listeners utilize these constraints to guide the interpretation?*

Integration is an important aspect of incremental speech comprehension which allows each word to be interpreted in the light of the contextual representation. This allows listeners to rapidly construct the sentence-level understanding as each word incrementally unfolds over time. By calculating the index of integration (see Chapter 2), I constructed the syntactic and semantic models which were tested after the onset of a target word in a sentence in order to address this question.

Chapter 4 is concerned more with the issues of incrementality in human speech processing and reliance on explicit syntactic rules in understanding speech. In this chapter, I use a state-of-art neural network model trained on large-scale corpora to predict an upcoming word as accurately as possible in a sentence (Jozefowicz et al., 2016). This lexical predictive machine allows to evaluate the lexicalist claim in more detail, by addressing the following questions:

4) *To what extent is human speech comprehension incremental? (or, more specifically, do these predictive computations occur for every word in a sentence?)*

This question is specifically concerned with the granularity in the level of predictive computations. The predictive machine naturally computes and updates the constraints at every word in a sentence as it is trained to do so. However, does human speech comprehension, whose goal is to understand the message that a speaker conveys, show the same level of predictive computations? Chapter 4 addresses this question by comparing the internal and output representations of this predictive machine at every point from a subject noun to a complement noun in a sentence with the representations of neural activity aligned to each of these points.

5) *Is it possible for a model, which learned statistical relations among different words through a large corpus but does not have any explicit knowledge of syntax, to explain human speech processing?*

It is important to highlight that this model does not have any explicit syntactic knowledge (the available syntactic knowledge in this model is only implicitly learned from the word-level statistics if it improves the accuracy of lexical prediction). By addressing this question, this chapter further illuminates how necessary the hierarchical processing is in understanding a sentence (with simple grammatical structures in daily conversation).

In order to address these questions, the EEG/MEG signals were recorded while participants were listening to natural sentences throughout the experiment.

1.6. Preparing EEG/MEG data for the investigation of speech comprehension

As mentioned above, EEG (and MEG) are the time-sensitive brain recording devices that preserve temporal dynamics of neural activity. Such “temporal dynamics” provide an essential source of variability to investigate various computations involved in incremental speech process in human brain. Further, with the developments of source-reconstruction techniques, the activity recorded at each electrode/sensor could be used to reconstruct the source activity inside the brain. Such source-reconstruction techniques provide useful spatial dynamics, allowing researchers to test their hypothesis regarding “where” in the brain the effect being modelled would occur (As a limitation, the original source activity can spread through MEG source estimation, leading to false positive interpretation of brain areas (Sato, Yamashita, Sato & Miyawaki, 2018) and zero-lag correlation among nearby sources (Colclough, Brookes, Smith & Woolrich, 2015)). This section aims to explain how the EEG/MEG data that are used throughout this dissertation are recorded, processed and source-reconstructed to investigate the spatiotemporal patterns of neural activity.

1.6.1. Electro- and Magneto-encephalography

Both EEG and MEG are the devices which record the time-varying electrophysiological activity in the brain. More specifically, they record signals from the post-synaptic potential (PSP) occurring at the dendrites of the pyramidal cells. These cells are one of the principal cortical neurons which lie perpendicular to the cortical surface. Depending on the post-synaptic receptors activated by neurotransmitters released from the pre-synaptic axon terminal, the flow of charge changes the membrane potential in the post-synaptic dendrites. Therefore, in contrast to fMRI which records the depletion of oxygen in blood flow in the brain as a measure of the neural activity (hemodynamics), EMEG captures more direct electrophysiological dynamics without losing the temporal resolution. These devices are blind, however, to electric currents generated by the action potentials propagating along the axon because the currents of opposite polarity always flow in vicinity rendering the action potentials invisible.

Each EEG sensor measures the voltage fluctuation on the scalp generated by the substantial number of charges at the apical dendrites during the post-synaptic potential. The net voltage on the scalp depends on the tissue conductivity of the brain, skull and scalp, and the distance between the charge and the scalp (assuming the homogeneity of tissue conductivity, the distance is often a more influential factor than the conductivity in practice). On the other hand, each MEG sensor measures the strength of magnetic fields generated by electric currents. A magnetic field is always perpendicular to the direction of the electric current. As a result, a common view is that MEG picks up the source activity, oriented tangentially to the scalp because the magnetic fields perpendicular to those neurons' orientation is always parallel to the MEG sensor directly above it. However, such view is held only for a spherical head model and only few cortical sources are exactly tangential in practice (Hillebrand & Barnes, 2002; Ahlfors, Han, Belliveau & Hamalainen, 2010). Rather, depth is more important factor that determines the detectability of a cortical current given that the field strength at the sensors is inversely proportional to the cubed distance. In summary, both EEG and MEG sensors record the electrophysiological activity of neurons but only MEG recordings are much less sensitive to the deeper sources.

Unlike EEG, there are two different types of MEG sensors: magnetometers and gradiometers. A magnetometer consists of a single superconducting pick-up coil which induces an electric current proportional to the magnetic flux, the surface integral of the magnetic field passing through the coil. In the human brain, there are a number of sources varying in their strengths and orientations over time. The magnetometer recordings, thus, reflect the sum of the magnetic fields at the surface of a sensor coil at a specific time-point. In contrast, a gradiometer measures the difference between magnetic fields recorded by two pick-up coils attached in a twisted manner via summation. In other words, a gradiometer measures the spatial gradient over the unidimensional space or axis. Hence, if the two pick-up coils collect the same amount of flux, the induced current in each of the coil will cancel out, leading to zero gradient. The strength of the magnetic field at the input coil generated by the net induced current from the pick-up coil is measured by a SQUID (superconducting quantum interference device). It is an extremely sensitive device used in both magnetometers and gradiometers which is capable of measuring very subtle magnetic fields greater than 10^{-14} $Tesla.$ Not surprisingly, it has been suggested that combining both EEG and MEG yields the most accurate localization (Sharon, Hamalainen, Tootell, Halgren & Belliveau, 2007) and the maximal average precision (Henson, Mouchlianitis & Friston, 2009) due to their complementary sensitivities depending on the depth of source dipoles.

MEG data were recorded on a VectorView system (Elekta Neuromag, Helsinki, Finland). The MEG machine consisted of 102 patches and each patch contained a magnetometer and two planar gradiometers (i.e. two pick-up coils attached next to each other) in orthogonal directions, designed to measure the spatial gradient over the lateral surface of the brain. This particular configuration is very efficient as each sensor in the same location (patch) measures independent information. As the SQUID device is extremely sensitive, the recordings were carried out in a magnetically shielded room to prevent the neural signal from being contaminated by the external electromagnetic noise. In order to monitor head movement in the MEG helmet, five HPI (head positioning indicator) coils attached to the scalp recorded head position every 200ms. In conjunction with the MEG recordings, EEG signals were also recorded using an MEG compatible EEG cap (Easycap, Falk Minow Services, Herrching-Breitbrunn, Germany) with 70 electrodes, plus a set of external electrodes and a nose reference. Blinks and eye movements were recorded by EOG (electro-oculogram) placed above and beneath the left eye and beside the left and right outer canthi. Cardio-vascular effects were also recorded by ECG (Electro-cardiogram) attached to right shoulder blade and left torso. Then, the positions of the HPI coils and EEG electrodes were digitized relative to the three anatomical landmarks including nasion, left and right peri-auricular points. The signals were recorded with a sampling rate of 1kHz and any MEG signals below 0.03Hz were high-pass filtered.

1.6.2. Participants

Fifteen participants (7 female; average age: 24 years; range: 18-35 years) took part in the study. They were all native British English speakers and right-handed with normal hearing. Two participants were excluded from the analysis because one of them fell asleep during the experiment and the other one had poor quality EEG recordings due to small head-size.

Informed consent was obtained from all participants and the study was approved by the Cambridge Psychology Research Ethics Committee.

1.6.3. Stimuli and Procedure

While the brain activity of each participant was recorded using EMEG, they listened to 200 spoken sentences, consisting of 50 sets of four different types. Each sentence consisted of a subject noun phrase ("The experienced walker") followed by a main verb ("chose"). We manipulated the probability of the verb's complement both syntactically and semantically based on the verb's subcategorisation preferences and its selectional restrictions. The combination of the subject noun phrase and the verb ("The experienced walker chose") was repeated four times followed by one of two function words associated with a particular frame which was either highly preferred ("the" for the direct object frame) or less preferred ("to" with a infinitival frame). Similarly, the probability of the noun (or verb) following the function word also varied (see Figure 3-1). All function words and nouns were natural continuations of the verb; the stimuli contained no violations.

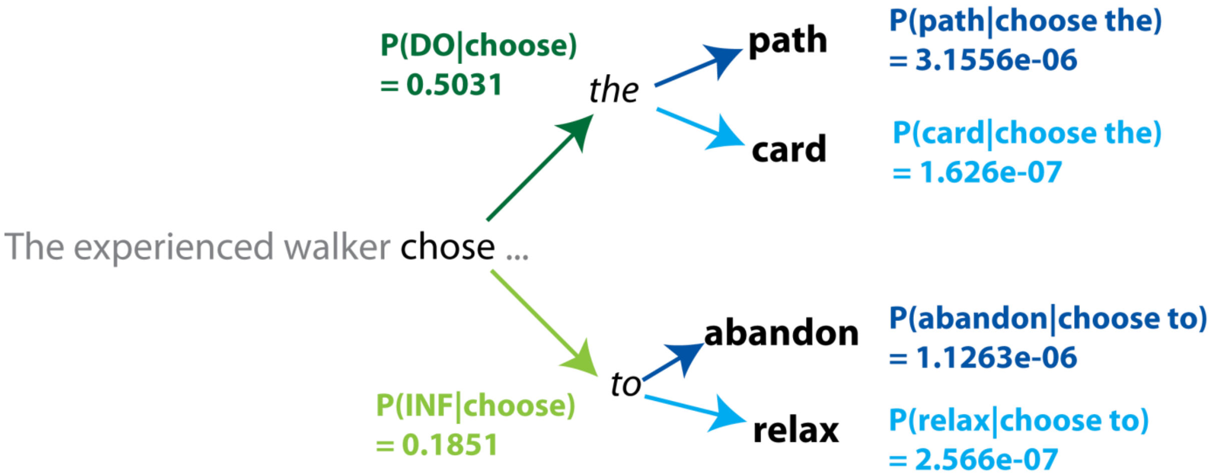


Figure 3-1: Design of the experimental stimuli. Each sentence contained a key main verb ("chose") followed by a complement function word ("the" or "to") that was either consistent with the verb's preferred subcategorisation frame (dark green) or with a less preferred frame (light green). A function word was followed by a noun or a verb that was either consistent with the verb's preferred continuation (dark blue) or with its less preferred continuation (light blue). This generated a set of four sentences for each context (i.e. subject noun phrase + verb) and there were 50 different contexts in total.

To construct these sentence sets, the main verbs were chosen from the VALEX database (Korhonen et al., 2006) that occurred with (at least) two different complement structures including a simple transitive direct object frame (e.g. "…chose the path…"). The other structure was one of three other possible structures including sentential complement ("…denied that the court…"), infinitival complement ("…wanted to become…") and prepositional phrase complement ("…fled to the forest…"). To ensure variability in the predictability of the complement nouns (or verbs), we varied the probability of these content words based on the preceding verb and the complement function word according to Google Books n-gram frequencies. In the end, 200 sentences, grouped into 50 sets of four, were constructed consisting of 100 direct object, 40 infinitival, 28 prepositional and 32 sentential complement structures with different complement content words. These sentences were spoken by a native female British English speaker and were recorded in a soundproof booth.

These stimuli were delivered to participants using MEG compatible earphones. Participants were asked to listen to the sentences attentively and were not given an explicit task to perform. The presentation order of the stimuli was pseudo-randomized and counter-balanced across participants. In each trial, a fixation cross was visually presented at the centre of the screen for 700ms followed by the spoken sentence stimulus then a silent inter-stimulus interval of 750ms and, finally, a blink break of 1000ms. Participants were requested to limit their blinking to this blink break period in order to minimize eye and body movement artefacts while listening to speech. Stimuli were presented using E-prime 2 (Psychology Software Tools).

1.6.4. EMEG pre-processing

During the recording session, the noisy EEG channels were identified and later removed. The initial pre-processing for the raw MEG data involved removing bad channels, compensating for head-movement by transformining the head position recorded by the HPI coils to a common head position and excluding any signals from outside the MEG helmet using signal space separation techniques (Taulu et al., 2005) using max-filter (Elekta-Neuromag).

Then, for both EMEG data, a low pass filter at 40Hz was applied using 5th order Butterworth Digital Filter in a zero-phase filtering framework using SPM8 (Statistical Parametric Mapping 8, Welcome Institute of Imaging Neuroscience, London, UK). In order to remove any physiologically driven artefacts such as blinks or cardiac signals recorded by EOG and

ECG, independent component analysis (ICA) was applied to the data. ICA is a widely used technique to decompose the data into a set of independent components (IC) either by maximizing the non-Gaussianity (mixture of components being more Gaussian than a single independent component) or by minimizing the mutual information between the components. Each IC was then correlated with the EOG and ECG channels using EEGLAB's infomax principle (Bell & Sejnowski, 1995; Delorme & Makeig, 2004). Any ICs showing very high temporal correlation (>0.3) with any of these channels were removed and the remaining ICs were visually inspected to ensure that no artefact component remained. The remaining ICs were used to reconstruct the data.

Next, five separate analysis epochs were generated for each trial by aligning the data to one of the three points of interest in each sentence (see Figure 3-3). After epoching, the data for each channel were baseline-corrected by subtracting the time-averaged data from a baseline period of -200ms to 0ms relative to the sentence onset (i.e. a period of silence immediately preceding the sentence). Finally, automatic artefact rejection was used to identify trials for which 15% or more sensors in any one of the three sensor types exceeded an amplitude threshold (6×10^{-11} T for magnetometers, 3×10^{-12} T/m for gradiometers and 2×10^{-4} V for EEG), and these trials (15 trials on average) were rejected. Any sensors that are consistently noisy and exceed the threshold for most of the trials were additionally marked as bad channels during visual inspection and removed from further analysis. These pre-processing steps were carried out using SPM8 (Statistical Parametric Mapping 8, Welcome Institute of Imaging Neuroscience, London, UK).

1.6.5. EMEG source reconstruction

Source reconstruction aims to estimate the regional response within a brain using the EMEG data recorded outside the scalp. For more accurate reconstruction specific to each subject's anatomical structure, structural MRI scans were acquired for each participant in a separate session using 1mm isotropic resolution T1-weighted MPRAGE on a Siemens 3T Prisma scanner (Siemens Medical Solutions, Camberley, UK). Participants' structural MRI images were first transformed into an MNI template brain which was then inverse-transformed to construct individual scalp and cortical meshes by warping canonical meshes of the MNI template to the original MRI space (Mattout et al., 2007). The MRI co-ordinates from individual scalp and cortical meshes were co-registered with the MEG sensor and EEG electrode co-ordinates using the digitized head-shape during data acquisition and aligning the digitized fiducial points to the fiducial landmarks defined on the subject's MRI image. A single-shell conductor model was used as a forward model for MEG recordings which assumes that all currents are generated inside the skull. For EEG forward modelling, we used a boundary element model (BEM) which defines three boundary layers (brain, skull and scalp) and assumes that the tissue conductivity inside each layer is homogenous. The forward modelling procedure computes the lead field matrix for each participant which defines the sensitivity of each source to each sensor (mapping matrix between sources and sensors). Although EEG, magnetometers and gradiometers were recorded in different measurement unit, they were effectively normalized by their respective average second-order moment (i.e. sample variance for the mean-corrected data). This procedure was similarly applied to the lead field matrix associated with each of different modalities. The normalized sensor recordings and lead-field matrices rendered different sensor modalities (and their associated hyperparameters of the error components) comparable and allowed them to be fused to yield a better precision of the source estimates than the precision from any of the unimodal inversions (Henson et al., 2009).

Given that the number of source dipoles is always greater than the number of sensors, there are an infinite number of solutions to estimating the source currents that generated the data. SPM source-reconstruction offers a Bayesian solution (a.k.a. Parametric Empirical Bayes) to this inverse problem, based on an assumption of source covariance as a prior (Friston et al., 2008; Lopez et al., 2014). Within this PEB framework, the source estimate $\hat{J}$ is expressed as the expected value of the posterior $E[P(J|Y)]$ conditioned on the sensor-level (multivariate) data Y. Assuming that J (true source activity) is a zero mean Gaussian process, the posterior $P(J|Y)$ can be formulated in terms of the multivariate Gaussian likelihood $P(Y|J) \sim N(LJ, Q_e)$ and prior $P(J) \sim N(0, Q_J)$ under Bayes' theorem (L = lead-field (sensors x source dipoles) matrix, Q_e = sensor noise covariance matrix, Q_J = source covariance matrix). Now, finding the estimate $\hat{J}$ can be simplified to maximizing $P(Y|J)P(J)$ given that $P(J|Y) \propto P(Y|J)P(J)$ and the remaining variables to be estimated are the two covariance matrices Q_e and Q_J (Lopez, Litvak, Espinosa, Friston & Barnes, 2014).

The noise covariance Q_e is typically in the form: $Q_e = h_0 I_{N_c}$ where h_0 is the sensor noise variance and I_{N_c} is a sensor x sensor identity matrix, reflecting that the sensor recording at each location is orthogonal and all sensors are affected by the same amount of noise variance.

This covariance parameter works as a regularization parameter in the framework (i.e. αI in Tikhonov regularization in the form $(A + \alpha I)x = b$), producing the regularized source estimate $\hat{J}$. In order to compute the optimal source covariance Q_J, another optimization objective is introduced in the framework to obtain the hyperparameter h that maximize the model evidence $P(Y) = P(Y|h)$. The source J is parameterized by h which determines the size of a prior variance in the source space. Then, the computation of this model evidence involves the data covariance matrix $\sum_Y$ which is composed of the noise covariance Q_e (error) and the projected source covariance onto the sensor space LQ_JL (signal). This projection renders the objective for h to be formulated exclusively in terms of the data, allows the regularization parameter to be treated as another hyperparameter during the optimization and makes the whole framework computationally feasible (see Lopez et al., 2014).

In this thesis, the source dipoles were assumed to be independent and to have equal variance (minimum norm assumption; Hamalainen & Ilmoniemi, 1994). This source prior was empirically adapted using the hyperparameter which was, in turn, used to compute the maximum a posterior (MAP) source estimate (Dale & Sereno, 1993). After the source reconstruction, the time-course of each source vertex was extracted for further analysis. In the next chapter, all computational models used to characterize the spatiotemporal patterns of the source-reconstructed EMEG data are discussed.

Chapter 2: Computational modelling of the incremental processing of a sentence

In this chapter, I describe all methodological details and motivations about the computational models I generated to investigate incremental speech processing in humans. Under the view that human speech processing is predictive (Kuperberg & Jaeger, 2016), I focus specifically on modelling the multi-level constraints and introduce the Bayesian Belief Updating (BBU) framework as a descriptive measure of incremental speech comprehension (Kuperberg, 2016). Then, I explain behavioural and computational approaches to modelling constraints using Cloze probability and neural network models in the connectionist framework. Delving into a number of network architectures and training algorithms, I motivate the use of recurrent network with a memory cell (long-, short-term memory (LSTM), Jozefowicz et al., 2016) consisting of a number of gate functions which determine the content to be preserved, forgotten and extracted and the adaptive training algorithms which enables the network to flexibly attend to the informative teaching materials (i.e. larger gradient). The softmax output distribution from this LSTM network was used as a model of lexical constraint based on the given context. Then, a model of update (or integration) and entropy (informativeness of the constraint) is derived from the constraint in the light of the information theory. The behaviour of these commonly used metrics (surprisal and entropy) is interpreted in relation to the constraint and motivated as a model of human cognition.

In the following sections, I describe the computational models of syntactic and semantic constraints. I used both the VALEX database from Korhonen, Krymolowski and Briscoe (2006) and the data collected from a continuation prêt-test for computational and behavioural modelling of the syntactic constraint. Also, I consider three different approaches to modelling semantic constraint: 1) propagating the lexical constraint to the pre-defined semantic space and 2) applying a dimensionality reduction technique to the lexical constraint and 3) training a Bayesian topic model in the LDA framework using the VB or the Gibbs sampling algorithms. Each of these approaches (and their derivations) is described in detail in comparisons with each other.

2.1. Bayesian Belief Updating (BBU)

Incremental speech processing involves using the available information from the context to constrain an upcoming input (which can be a word, a phrase, a sentence etc.) and integrate it into the prior context once it is heard in order to constrain a subsequent input more accurately. This cycle continues until the speaker ends his message. This conceptual description of incremental speech processing fits well in the Bayesian framework of language comprehension. The motivation of this framework originates from Bayes' theorem which describes the probability of an event based on the prior information and knowledge related to the event. A simple mathematical description of Bayes' theorem is as follows:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \dots (1)$$

where A is a target variable and B is a context variable on which the target A is conditioned on. As a simple application to language processing, suppose that a listener hears an adjective-noun phrase like "*yellow banana*". The goal is to model the listener's internal beliefs about "*banana*" given the preceding adjective "*yellow*". By simply substituting A with "*banana*" and B with "*yellow*", we obtain the following:

$$P("banana"_t|"yellow"_{t-1}) = \frac{P("yellow"_{t-1}|"banana"_t)P("banana"_t)}{P("yellow"_{t-1})} \dots (2)$$

where t and $t-1$ indicates the relative position of each word in the phrase. The goal is to model the posterior $P("banana"|"yellow")$ describing the probability of "*banana*" given "*yellow*". This expression already proves its usefulness by showing an explicit mapping between the goal (posterior) and the prior. The prior $P("banana")$ describes the listener's beliefs about the target "*banana*" (i.e. subjective probability of "*banana*" alone) before knowing the context "*yellow*". Then, the likelihood $P("yellow"|"banana")$ evaluates the context "*yellow*" against his prior beliefs about the target "*banana*". The evidence $P("yellow")$ works as a context normaliser whose practical role is explained in Footnote 1 in Chapter 1. The concept of belief updating is reflected by the shift from a prior to a posterior at any given cycle until the posterior converges to the delta distribution (target = 1 or 0 otherwise). In a modelling perspective, this Bayesian approach provides useful insight into how prediction may change and develop as new words are incrementally unfolded in a sentence.

Another important aspect of this approach is that it models the cyclical development of prediction in sentence and discourse comprehension. Suppose that we are modelling the listener's syntactic prediction of a complement structure in a sentence: "*The intrepid child found the picture*". For illustration purposes, I assume that the subject NP "*The intrepid child*" is independent of the following complement structure such that it is constrained entirely by the verb "*found*" in a preceding context. Then, it is possible to track changes in prediction as follows (Figure 2-1):

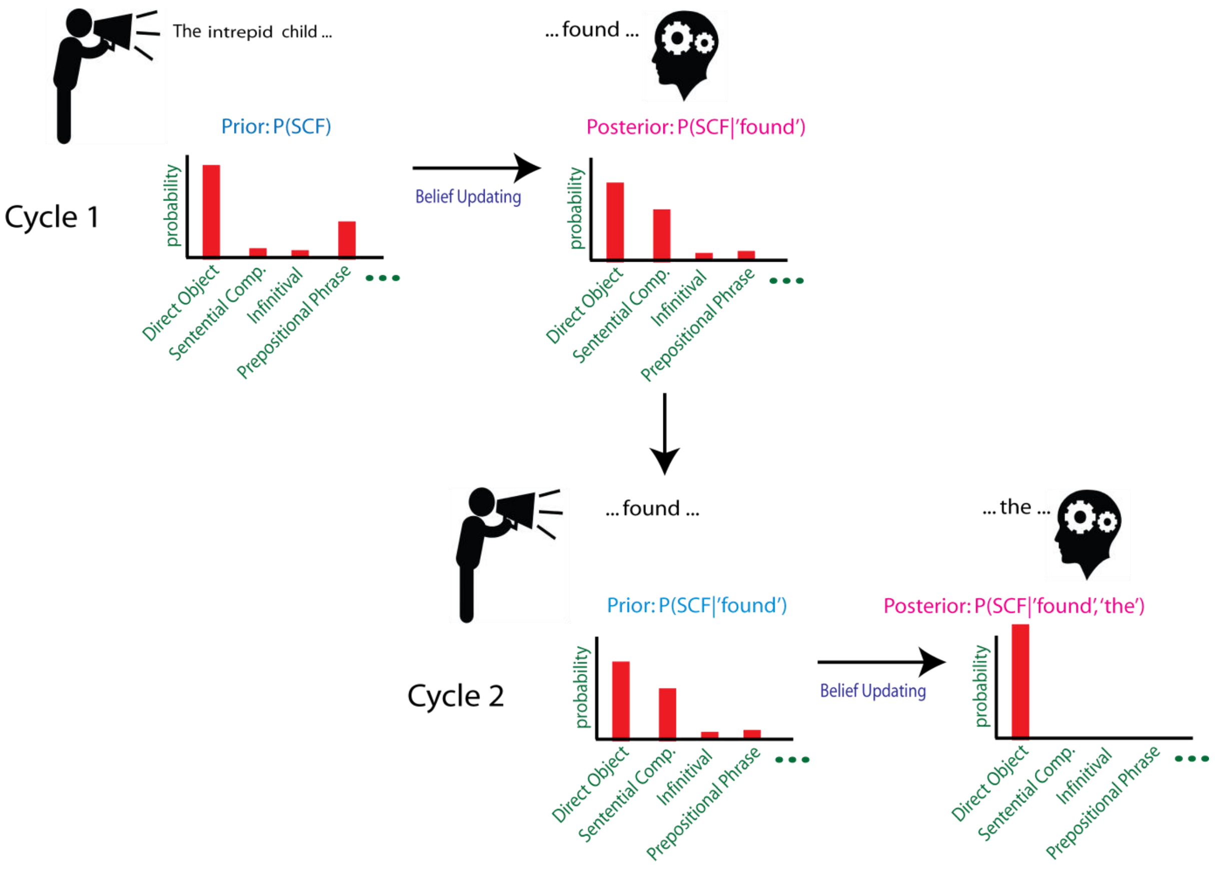


Figure 2-1: A simplistic visual illustration of belief updating about the complement syntactic structure across different cycles in time. SCF = subcategorization frame.

In Figure 2-1, Cycle 1 describes the process of incorporating the main verb "*found*" into prediction. Cycle 2 shows that this verb-incorporated prediction becomes a new prior to constrain the syntactic frames. As a direct object structure is confirmed by the determiner "*the*", the prediction cycle ends in Cycle 2 in this example and the prior facilitates the integration of the direct object structure into the sentence. Hence, by tailoring the prediction more specifically to the up-to-date context, this Bayesian model promotes more rapid and accurate integration of the target frame (direct object). It is worth noting that any posterior at the end cycle (Cycle 2 in this example) converges to a delta distribution and the process of belief updating becomes conceptually equivalent to integrating the target into the context (the "target", in practice, refers to a specific property (e.g. semantic meaning or grammatical category etc.) of a particular linguistic unit (e.g. a word, a phrase, a clause etc.) that appears after the context).

As shown in (2) and Figure 2-1, incremental speech comprehension proceeds with updating the beliefs each time an input (i.e. verb) that constrains the target (i.e. SCF) is heard. However, as already discussed in Chapter 1, prediction in speech processing is not merely limited to words but includes a variety of linguistic aspects from perception (phonological-lexical) to cognition (syntax-semantics). The psycholinguistic accounts based on the Fodorian modular theory (Fodor, 1983) claims that the processing streams are organized into separate, autonomous modules (Frazier, 1987). Other accounts propose jointly interacting streams (Marslen-Wilson, 1975; Altmann & Steedman, 1988). In this section, I briefly review a recent generative framework proposed by Kuperberg (2016) in the Bayesian perspective.

Kuperberg's framework claims that listeners infer the underlying cause of the observed inputs from a set of hierarchically organized representations (or internal generative model). These representations best explain the statistical properties of the observed inputs based on their beliefs about the message that the speaker tries to convey. The beliefs propagate down to lower levels to tailor the representations by generating probabilistic predictions before processing the new input. Predictions at these various domains hierarchically interact with each other: for example, predictions about semantic meanings or syntactic structures of possible continuations could influence the predictions about candidate words which could, in turn, affect the expected sequences of phonemes. These probabilistic predictions are evaluated against the bottom-up evidence once the new input is heard to update their prior beliefs. This top-down prediction scheme facilitates the processing of an input word in a sentence and the input, in turn, enables flexible updating of the multi-level constraints through bottom-up projections. This process is simplistically illustrated in Figure 2-2 below.

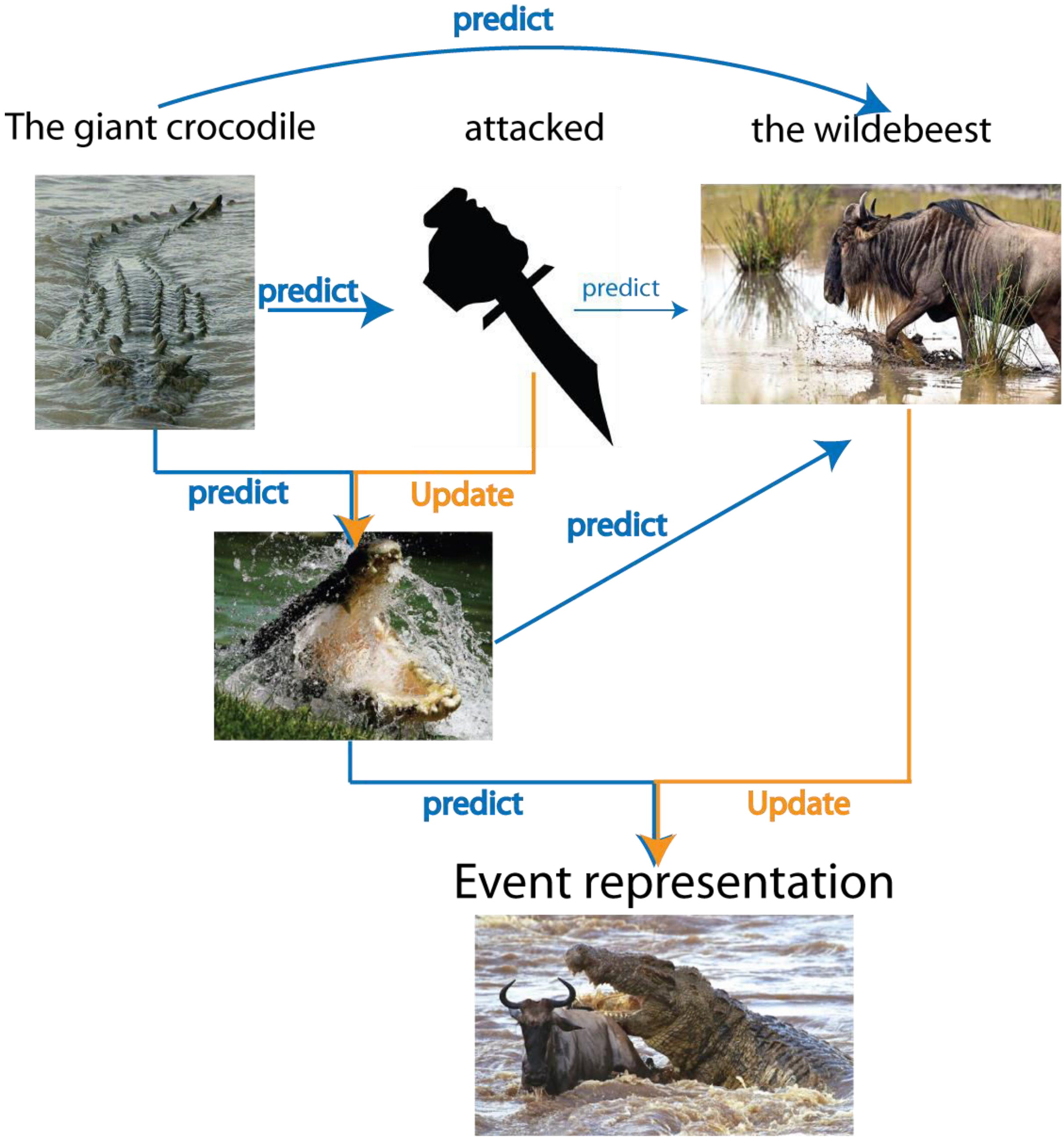


Figure 2-2: Incremental speech processing of a simple direct object sentence "The giant crocodile attacked the wildebeest" in the light of the BBU generative framework (Kuperberg, 2016). This describes the role played by each input (i.e. a subject noun phrase, a verb and a complement noun phrase) in constructing the event representation (i.e. a message) in a predictive processing framework. Blue arrows indicate "prediction" and orange arrows indicate "update" or "integration".

Now, the problem simplifies to characterizing the arrows in Figure 2-2: prediction and update. Under the view of prediction as a graded/probabilistic phenomenon (see Kuperberg & Jaeger, 2016), the conditional probability distribution about the upcoming input directly represents information used to predict the upcoming input (i.e. constraints). Also, it is important to quantify the certainty of beliefs because the strength of top-down prediction depends on the certainty with which the beliefs are held (Kuperberg, 2016). Lastly, the difficulty of updating reflects the proportion of variance in constraints (a.k.a. "pruned probability mass" in Levy (2008, p. 1131)) which cannot be explained by the bottom-up input, so-called "prediction error". The human language system aims to minimize this prediction error by an iterative process of predicting and updating throughout a sentence and will eventually obtain converged representations at various levels each of which best explains the observed sentence. The ways to characterize prediction and to quantify certainty and error are described in the following sections.

This Kuperberg's BBU framework is a variant of "predictive coding" framework (Friston, 2005, 2008) which has drawn significant attention in the field of cognitive/perceptual neuroscience. As stated in Kuperberg and Jaeger (2016), *"Hierarchical predictive coding in the brain takes the principles of the hierarchical generative framework to an extreme by proposing that the flow of bottom-up information from primary sensory cortices to higher level association cortices constitutes only the prediction error, that is, only information that has not already been "explained away" by predictions that have propagated down from higher level cortices…"*. This specific neurobiological hypothesis from the predictive coding account has been tested and corroborated in a series of behavioural and neuroimaging studies of speech perception (Sohoglu, Peelle, Carlyon & Davis, 2012, 2014; Sohoglu & Davis, 2016). They consistently reported the reduced activity in superior temporal gyrus (STG) when the speech input (target) was more expected, supporting the claim that brain is sensitive to the mismatch (error) between expected and actual input.

2.2. Computational and behavioural modelling of prediction

The most straightforward approach to model human prediction is to ask individuals directly what they predicted in a given context. By asking many individuals, it is possible to count the total number of individuals who predicted an item for all available items. Normalising by the total number of counts across all items gives a probability distribution that directly represents human constraint (also known as Cloze probability; Taylor, 1953). More generally, these behavioural responses reflect the ‘maximal incremental interpretation’ of the context (Marslen-Wilson, 1975; Tyler & Marslen-Wilson, 1977; Marslen-Wilson et al., 1993) – namely, the integration of the lexical syntactic and semantic information carried by the words heard so far into an interpretation of the utterance fragment in terms of the listener’s knowledge of the world and likely event structures in the context of that knowledge. For example, in the context like “if you walk too near to the runway …”, the on-line choice of the adjectival interpretation of the subsequent phrase “landing planes” (Tyler & Marslen-Wilson, 1977) reflects both the lexical syntactic and semantic properties of the words in this phrase and pragmatic inference operating over the listener’s knowledge of runways, relative distance from the runway, the properties of landing planes, and so forth. Similar wide-ranging processes are expected to be operating in the incremental interpretation of the subject noun phrase (SNP) + verb (V) contexts. In summary, this approach captures listeners’ subjective expectation about potential candidates with the varying degree of preference in psycholinguistic modelling. Previous ERP studies have used this approach and showed that N400 amplitude decreases for items with higher Cloze probability (Delong et al., 2005; Federmeier et al., 2007).

In contrast, another approach extracts the probability from the frequency of every item in a corpus. A corpus is a large text database processed and stored from one or more sources such as books, newspapers, broadcasts etc. With the large amount of data in the corpus, it is possible to obtain a very accurate, objective probability distribution. Therefore, models of constraints constructed from the behavioural data are inevitably based on a much lower number of samples than the corpus-based constraints models. Further, the corpus-based constraints are free from any non-linguistic variables such as recent experience or metacognitive strategies that may affect subjective expectations as in Cloze probability. Assuming that the mapping function between the observed (objective) and the perceived (subjective) probability is approximately identity (Gallistel et al., 2014), psycholinguistic application of the corpus-based probabilities for modelling human prediction is motivated. However, the obvious limitation of corpus-based approach is that the total number of unique samples in the corpus must grow exponentially with every word being added to the context due to the combinatorial explosion of linguistic contexts.

An alternative approach of modelling a predictive process during speech comprehension based on the connectionist view captures the important properties in the entire context and utilizes them to generate an accurate constraint. Consequently, this approach provides a system that generates an output from its internal state that has been altered by a current input, instead of directly retrieving from a lexical database. The way that the state is altered is based on the previous experience from training.

In this thesis, I use each of these approaches to address different questions. First, I consider the corpus-based approach for modelling lexically-driven (verb-based) constraints since it has been shown that a verb provides multiple levels of predictive information (see Trueswell et al., 1993; Gibson & Pearlmutter, 1998; Bicknell et al., 2010; Elman, 2011). Using this model, I aim to address if such lexically-driven constraints are relevant in sentence processing that often contains multiple words in a context. Second, I construct models of constraints based on behavioural data to investigate the facilitatory role of both syntactic and semantic constraints based on the entire context in processing the upcoming input (i.e. verb's complement). Lastly, using the models of constraints based on the connectionist view, I ask to what extent the listeners' predictive processing is incremental during speech comprehension. By addressing this question, I aim to elucidate the level of specificity in predictive processing during incremental speech comprehension in the brain (see Kuperberg, 2016 and Figure 2-2). In the next section, I describe the architectures and training algorithms that are used to train connectionist models and evaluate them in the light of this experimental question.

2.3. Modelling prediction with neural networks

Owing to technological developments, many variations of connectionist models have attracted attention from many interdisciplinary researchers and various industries. They are known as neural networks, designed to perform particular tasks. In language modelling, they are typically trained to generate likely words (or other linguistic units) based on the given context. By inquiring about the likely upcoming words at every word in a sentence, it is possible to model the incremental development of prediction with these optimized connectionist machines. In this section, I describe important basics of neural networks and apply them to delve into a number of variations in the neural networks for language modelling.

2.3.1. Capturing non-linear patterns in the data using non-linear functions of linear classification algorithms

A neural network is a biologically inspired information processing system consisting of densely interconnected nodes (neurons) which are trained to solve specific problems. They have become the most successful and popular algorithms in the fields of data mining and machine learning due to their ability to learn complex non-linear patterns that exist in the data. In fact, finding a non-linear pattern is an appealing trait that distinguishes it from other widely used linear pattern classification algorithms such as logistic regression or support vector machine (SVM). In its simplest form, there are three layers including input, hidden and output layers, each of which consists of a set of neurons illustrated in a figure below (Figure 2-3).

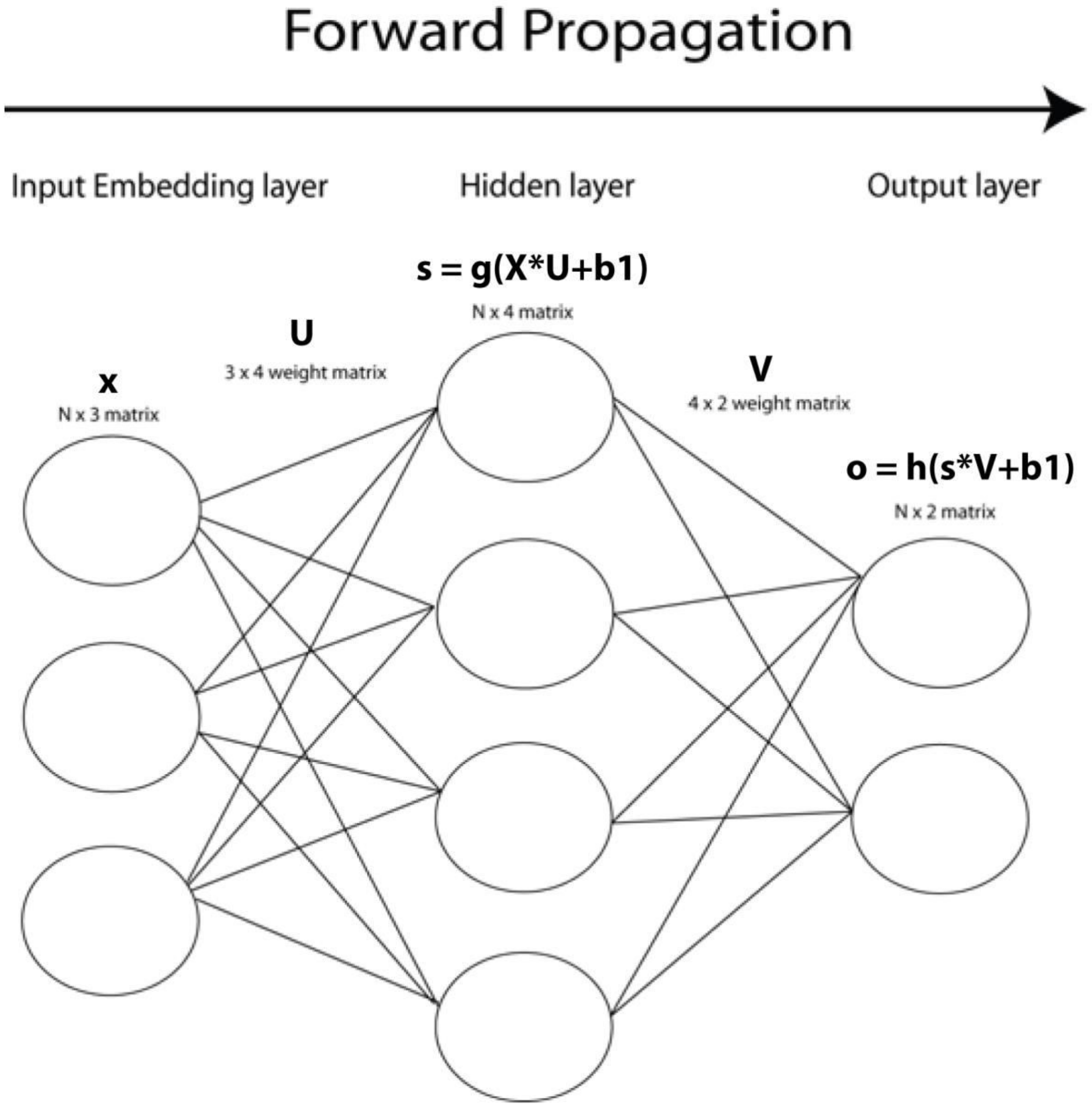


Figure 2-3: Visual illustration of architecture of a simple feed-forward neural network. x is a matrix of input embeddings, s is a matrix showing a hidden layer state and o is a matrix of an output. g and h are some non-linear functions and b1 and b2 are bias parameters. N is the total number of (batch) samples in the data.

From Figure 2-3, suppose we remove the hidden layer from its architecture and send the input directly to the output layer. With a particular function h, the neural network simply becomes equivalent to some well-known linear classification algorithms such as logistic regression with h being sigmoid, SVM with h being rectified linear unit (ReLU) and multinomial logistic regression with h being softmax. However, with the hidden layer intercepting the input in between, the linear combination of input features (also called predictors or independent variables) is non-linearly transformed by the function g which is, in turn, projected to the output layer. This effectively allows the algorithm to find a non-linear instead of a linear decision boundary. For this reason, the function g MUST be a non-linear function (regardless of how many times the function is applied, the output is still linear to the input if the function is linear).

In practice, there are two major non-linear functions that are commonly employed:

$$Sigmoid(XW1 + b1) = \frac{1}{1 + e^{-(xU+b1)}} = \frac{e^{xU+b1}}{e^{xU+b1} + 1} \dots (3)$$

$$ReLU(XW1 + b1) = \max(0, xU + b1) \dots (4)$$

(3) is a sigmoid function, used in logistic regression to generate a classifier response from the linear combination of input features. Using this sigmoid function, logistic regression models a log-odds of the binary response based on a linear combination of the input features. It is worth noting that this sigmoid function is a cumulative distribution function (CDF) of a normal distribution. This is one of its most appealing traits as a classifier function given that evidence is accumulative in real-life decision-making. For example, the grey sky makes people’s expectation of the rain even stronger after watching the weather forecast predicting the rain. Hence, modelling their responses (whether to bring an umbrella or not) should accumulate the evidence over the number of input features (e.g. grey sky, weather forecast etc…) and return “bring umbrella” if the accumulated evidence exceeds the probability of 0.5 for rain. Any negative input value to this function returns an output value lower than 0.5

whereas it returns an output value higher than 0.5 with any positive input value (see Figure 2-4).

(4) is a rectified linear unit, used in SVM for the same purpose. SVM is a geometrically motivated classification algorithm which finds the optimal decision boundary by maximising the distance from it to the nearest data-point on each side as well as minimizing the classification error. This function always returns zero if the input value is negative or an output value above zero if the input value is positive (see Figure 2-4). There are two unique characteristics that render this function particularly attractive over the others. Due to the inherent sparseness (or unsmoothed representation of non-linearity) of this function, it is computationally efficient. A dense (or smoothed) representation is sensitive to any changes in the input whereas ReLU clearly distinguishes the inputs which are able to affect the representation from which aren't. However, as a side effect, this raised an issue of having dead neurons in the network (i.e. some nodes are not active whatsoever) being plunged into a perpetually inactive state. Another important characteristic is that its derivative is binarized into zero and one. It contrasts with the derivative of a sigmoid which is always in a range between zero and one (see Appendix 6). Consequently, a large network with multiple layers having sigmoid as an activation function suffers from the notorious "vanishing gradient" problem (i.e. If the input value is extreme OR if the network has many hidden layers, the sigmoid gradient quickly becomes zero due to the multiplicative nature of learning through backpropagation (see Appendix 7) whereas ReLU is immune to this problem.

There are variants of these functions (3) and (4) which are also commonly used: the hyperbolic tangent (rescaled sigmoid) and softplus (smoothed ReLU whose derivative is sigmoid; see Appendix 6) but they are beyond the scope of my thesis (see Figure 2-4 for visual illustration)

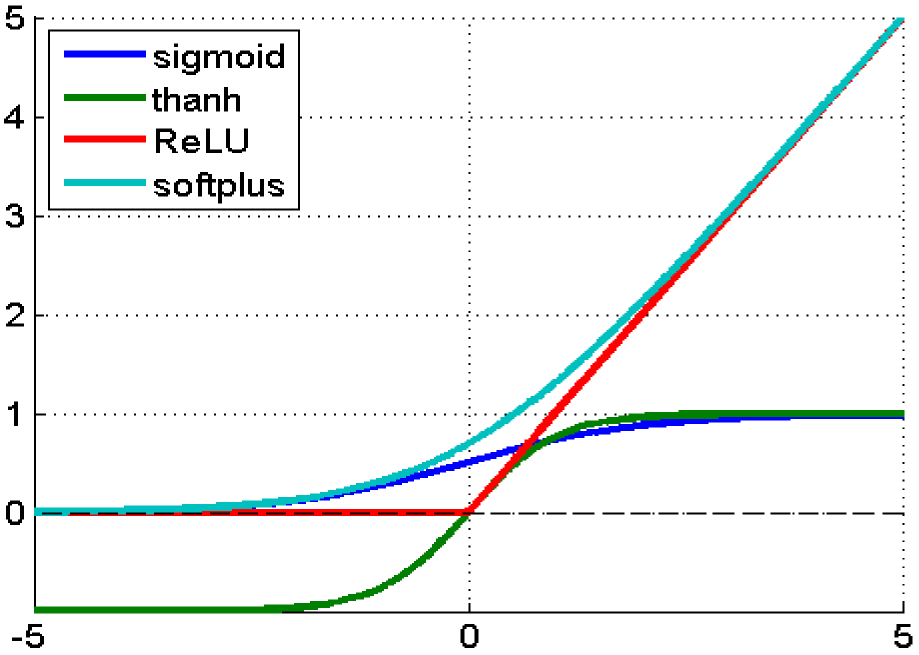


Figure 2-4: A graphical comparison of the common non-linear activation functions

2.3.2. Output layer and softmax

Softmax is the most commonly used activation function in the output layer of a neural network. It can simply be viewed as an exponential probability function to an input variable:

$$softmax(\theta)_j = \frac{e^{\theta_j}}{\sum_{i=1}^{No} e^{\theta_i}} \dots (5)$$

where No is a total number of output units. This function is particularly attractive because 1) it is a differentiable function that nicely translates the input values to a normalized scale and 2) it is a multi-class generalization of the logistic function which is designed to model a multinomial response variable.

2.3.3. Further implementation details

Further technical details of the neural network training are clearly explained in detail in Appendix 7 including the mathematical derivations of backpropagation and gradient optimization algorithms. The last paragraph of this section briefly discusses the practical viability of different batch sample training methods and describes how to treat the samples for an efficient optimization.

Adaptive optimizers

In this section, I briefly describe the actual optimization algorithm used to train the LSTM model in this thesis (i.e. Adaptive Gradient).

The optimization algorithms used in practice are more elegant variants which flexibly vary a learning rate η instead of setting it as a fixed parameter. Training data is often very sparse and various features occur in different frequencies, especially in natural language processing (NLP). Sometimes, infrequently occurring features are highly informative and, therefore, optimization can be greatly enhanced by pre-emphasising them. ADAGRAD (Adaptive Gradient; Duchi et al., 2011) is a variant of the gradient decent which applies η more flexibly depending on the previous error gradients up to the current update. By setting $g2_{qj}(t) = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial W2_{qj}}H(Y,O)_t$ and $g1_{pq}(t) = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial W1_{pq}}H(Y,O)_t$ where t represents the time of the current update (see Appendix 7), the ADAGRAD algorithm can be expressed as:

$$W2_{qj} \coloneqq W2_{qj} - \frac{\eta}{\sqrt{\sum_{\tau=1}^{t} g2_{qj}(\tau)^2 + \epsilon}} g2_{qj}(t) \ldots (6)$$

$$W1_{pq} \coloneqq W1_{pq} - \frac{\eta}{\sqrt{\sum_{\tau=1}^{t} g1_{pq}(\tau)^2 + \epsilon}} g1_{pq}(t) \ldots (7)$$

where ϵ is a smoothing term. The learning rate η at time t is adapted by the squared sum of past gradients with respect to a particular connection weight. This suggests that the update will be greater if the squared sum is low possibly because 1) not much error has been made by the network so far or 2) the neurons associated with the connection have not been responsible for the error as much as the others. Therefore, when these neurons that have rarely activated (thus, less responsible for the error so far) activate strongly at current update time t, the learning rate η gets relatively larger, naturally attracting the algorithm to attend to the connection between them. This naturally leads to the interpretation of pre-emphasising the infrequently occurring features associated with these rarely activated neurons. Note that the network model used in this thesis was trained from this ADAGRAD algorithm. Despite these benefits it brings to the optimization, it still suffers from a problem: shrinking learning rate (this will be revisited in discussion in Chapter 4).

2.3.4. Adding recurrence in the network

Although a simple neural network can be trained to generate an accurate prediction based on the given linguistic context, it lacks one of the most important aspects of human speech processing. Speech comprehension in humans involves understanding the relationship between sequentially unfolding words over time and interpreting them in the context of each other. The cognitive significance of "time" is not merely limited to language as human behaviours are generally co-ordinated in time. It directly implies causation and understanding the causal relationship between the series of behaviours over time, in turn, enlightens one's metacognitive processes. Therefore, any plausible cognitive model of human behaviours must represent temporal relation between the sequences of events.

An intuitive approach is to express time explicitly as an input in a form of a vector (or matrix). The first element in this vector represents the first temporal event, the second element represents the second temporal event and so on. However, the duration (or the number) of events often vary in practice and such events cannot be compared in this framework (i.e. all vectors must be in same length). Also, consider the following two vectors:

$$[0\ 1\ 1\ 1\ 0\ 0\ 0\ 0\ 0]$$

$$[0\ 0\ 0\ 1\ 1\ 1\ 0\ 0\ 0]$$

Although these vectors could plausibly reflect the same basic pattern in time (e.g. "He ***chose the path*** that ran by the river" vs. "The experienced walker ***chose the path*** crossing the river"), they can be judged as highly dissimilar because of the geometric difference in their absolute temporal positions (Elman, 1990). Rather than providing the information about time explicitly as an input in a specific format, Elman (1990) argued for representing time implicitly by its effects on processing. In this perspective, an input is an operator on the mental state such that it alters the state of the system to produce a goal-oriented behaviour. Then, the implicit representation of time can be expressed by adding recurrent links between the states of the system over time (see Figure 2-5).

Neural networks with these recurrent links are called recurrent neural networks which are common approach for language modelling in these days. Unlike a simple neural network whose prediction is purely based on the current input, a recurrent network alters the previous internal state based on the current input (see Figure 2-5).

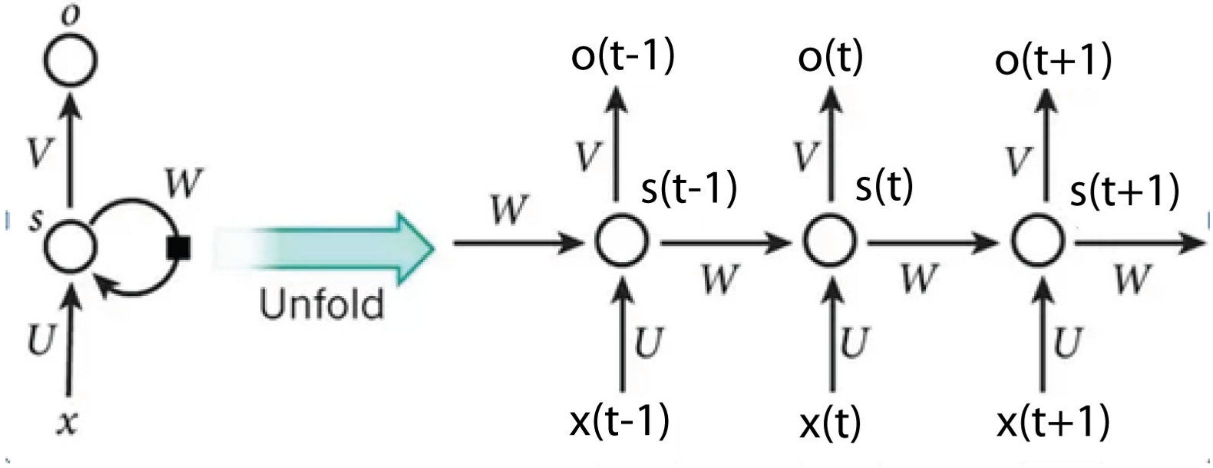


Figure 2-5: Visual illustration of a recurrent neural network. x, s *and* o *are input, hidden and output representations respectively.* U *is a weight matrix that projects the input* o *at any arbitrary given time* t *to the hidden layer* s *at* t. W *is a weight matrix mapping the previous hidden state* $s(t-1)$ *to the current state* $s(t)$ *(i.e. a recurrent link).* V *is a weight matrix mapping the hidden state* s *to the output* o. *Note that the recurrent link* W *is a new feature added to this recurrent architecture that does not exist in a simple neural network in Figure 2-3. With this addition, the concept of "time" is now implicitly represented by the architecture.*

The forward propagation in this architecture can be expressed by a set of equations below:

$$s(t) = \sigma(x(t)U + s(t-1)W + b1) \ldots (8)$$

$$o(t) = \varphi(s(t)V + b2) \ldots (9)$$

where σ and φ are the arbitrary non-linear activation functions at hidden (e.g. sigmoid) and output (e.g. softmax) layers respectively and $b1$ and $b2$ are the bias terms, allowing the layers to model the data space centred on some point other than the origin. Other notations are as described in Figure 2-5. Without the $s(t-1)W$ term in (8), the propagation becomes exactly same as a simple feedforward neural network described above.

Training RNN works similarly to a simple neural network except that the recurrent link W is also trained by back-propagating the error gradient through time using the chain rule as described in Appendix 7:

$$\frac{\partial}{\partial W_{q1,q2}} H\big(Y(t), O(t)\big) = \sum_{j=1}^{J} \frac{\partial H\big(Y(t), O(t)\big)}{\partial s1(t)_j} \frac{\partial s1(t)_j}{\partial s(t)_{q2}} \frac{\partial s(t)_{q2}}{\partial s2(t)_{q2}} \frac{\partial s2(t)_{q2}}{\partial W_{q1,q2}} \ldots (10)$$

where $s1(t) = s(t)V + b2$ and $s2(t) = x(t)U + s(t-1)W + b1$. Then,

$$\frac{\partial}{\partial W_{q1,q2}} H\big(Y(t), O(t)\big) = \sum_{j=1}^{J} V(t)_{q2,j}\big(o(t)_j - y(t)_j\big)s(t)_{q2}\big(1 - s(t)_{q2}\big)s(t-1)_{q1} \ldots (11)$$

This network only allows one adjacent previous state in time to influence the output. However, in a simple sentence "*The business owner declared bankruptcy*", the model will perform much better in predicting "*bankruptcy*" when it knows the subject "*The business owner*" on top of the verb "*declared*". In order to incorporate the contributions from every hidden state over time, it is necessary to sum up the contributions of each time step to the gradient. Following on from (10), it can be formulated as below:

$$\frac{\partial}{\partial W_{q1,q2}} H\big(Y(t), O(t)\big) = \sum_{j=1}^{J} \frac{\partial H\big(Y(t), O(t)\big)}{\partial s1(t)_j} \frac{\partial s1(t)_j}{\partial s(t)_{q2}} \sum_{\tau=0}^{t-1} \frac{\partial s(t)_{q2}}{\partial s2(t-\tau)_{q2}} \frac{\partial s2(t-\tau)_{q2}}{\partial W_{q1,q2}} \ldots (12)$$

Note that $\frac{\partial s(t)_{q2}}{\partial s2(t-\tau)_{q2}}$ can be expanded using the chain rule depending on τ. Hence, the error propagation through time can be computed by the extended formulation of (12). This is known as the back-propagation through time (BPTT) algorithm (due to the fact that the training becomes very difficult as $t \rightarrow \infty$, a practical implementation of BPTT back-propagates the error gradient only up to a certain time).

Not surprisingly, a recurrent neural network (RNN) generally performs better than the simple neural network when the inputs are sequences (like a sentence in language) instead of unrelated individual events. However, an important limitation of RNN is that it often fails to capture the long distance dependencies (e.g. the dependency relation between "child" and "smiled" in "The child who I thought you liked smiled"). This is mainly because of the “vanishing gradient” problem during training described above: with the derivative of sigmoid being less than 1 (i.e. ≤ 0.25), propagating the error through a number of recurrent layers necessarily forces the gradient to vanish (i.e. very close to zero), given the number of multiplications. One solution I suggested above is to use the ReLU instead of the sigmoid as its derivative is either 0 or 1 but this function brings other problems like dead neurons (i.e. a group of neurons can be plunged into a perpetually inactive state). To address this issue of vanishing gradient more effectively, a more sophisticated architecture called long short-term memory (LSTM) was introduced (Hochreiter & Schmidhuber, 1997).

2.3.5 Incremental language processing in a LSTM neural network

An LSTM network is a more sophisticated version of RNN which preserves the benefits of RNN as a model of incremental speech comprehension and additionally captures the long distance dependencies. In language modelling, LSTM is one of the most commonly adopted architectures for data mining and network training. Recently, Google announced a LSTM network trained on a 1 billion word benchmark which generates an accurate prediction of a following word based on the given context in a sentence (Jozefowicz et al., 2016). Note that the neural network model used in this thesis refers to this LSTM model. Here, I briefly walk through the architecture of LSTM (see also, Gers & Schmidhuber, 2000; Sundermeyer et al., 2015) and explain how it solves the vanishing gradient problem.

Instead of having a single operation in the recurrent hidden layer as in RNN, LSTM performs multiple operations, deciding which information to preserve and add inside the hidden layer. A useful analogy of this LSTM hidden layer is a memory cell with three gates in order to input, forget and output the contents of memory. First of all, it decides what to forget from the previous memory using the sigmoid function. Recall that the sigmoid function outputs a value between 0 and 1 which can be interpreted as a weight determining the strength of projection among the operators (a.k.a. gates in this analogy). Then, the vector of weights $\emptyset(t)$ reflects the state of the forget gate in the memory cell at a particular time t:

$$\emptyset(t) = \sigma(x(t)W_{x\emptyset} + s(t-1)W_{s\emptyset} + c(t-1)W_{c\emptyset} + b_{\emptyset}) \dots (13)$$

where σ is a sigmoid function, $x(t)$ is a current input with associated weights $W_{x\emptyset}$, $s(t-1)$ is a previous state in the hidden layer with associated weights $W_{s\emptyset}$ and $c(t-1)$ is a previous state in the memory cell with associated weights $W_{c\emptyset}$. Note that the cell state term $c(t-$

$1)W_{c\emptyset}$ does not exist in the RNN architecture. Again, this vector of the forget gate state $\emptyset(t)$ directly manipulates the memory content by setting 0 if it needs to be completely forgotten or setting 1 if it needs to be fully remembered.

Next, the LSTM network decides which information to add from the input and to store in the memory using sigmoid. With the same logic as above, the state of the input gate $\theta(t)$ can be expressed as:

$$\theta(t) = \sigma(x(t)W_{x\theta} + s(t-1)W_{s\theta} + c(t-1)W_{c\theta} + b_\theta) \dots (14)$$

Note that the weights to be trained in the input gate are different from those in the forget gate. From these weights that decide which memory contents to preserve from the previous cell state (or memory) $\emptyset(t)$ and that decide which information to store from the current input $\theta(t)$, we can construct new memory contents as below:

$$c(t) = c(t-1) \circledast \emptyset(t) + \tanh(x(t)W_{xc} + s(t-1)W_{sc} + b_c) \circledast \theta(t) \dots (15)$$

where $\tanh$ is a hyperbolic tangent function described in 2.3.1 and $\circledast$ denotes an element-wise product. Recall that $\tanh$ is a rescaled version of sigmoid in a scale between -1 and 1. Therefore, the input activation in the current hidden layer before passing through the memory cell is constructed through $\tanh$ which is, then, modified by the state of the input gate $\theta(t)$. Also, note that the element-wise product $\circledast$ allows a weight (a gate neuron in the input and forget gates) to directly modify a particular feature (either from the previous memory content or from the current input) processed by the neuron via one-to-one mapping (since a number of neurons in each gate in the memory cell is same). In summary, (15) shows that the modified input representation at the input gate is combined with the modified memory representation in the forget gate to generate a new memory content.

Lastly, the network decides what it is going to output. Similar to the state of the other gates, the state of the output gate directly modulates the new memory content from (15) using sigmoid:

$$\omega(t) = \sigma(x(t)W_{x\omega} + s(t-1)W_{s\omega} + c(t)W_{c\omega} + b_\omega) \dots (16)$$

These weights are used to modify the current memory content that is going to be output:

$$s(t) = \omega(t) \circledast \tanh\big(c(t)\big) \dots (17)$$

Similar to above, the unfiltered version of the memory content at the output gate is constructed through tanh which is, then, weighted by the state of the output gate through one-to-one mapping within every neuron in the output gate. Note that the bias term is not needed inside tanh of (17) because every distinct term that consists of new memory content $c(t)$ is already adjusted; see (13), (14) and (15). The gate response $s(t)$ (equivalent to the hidden layer activation in RNN) is then projected to the output layer of the network as in RNN (see (9)):

$$o(t) = \varphi\big(s(t)W_{go} + b2\big) \dots (18)$$

where φ is the softmax function to generate a probabilistic response. Then, the BPTT algorithm can be applied for optimizing every weight matrix (12 in total) through the memory cell from (13) to (18); see Figure 2-6 for illustration.

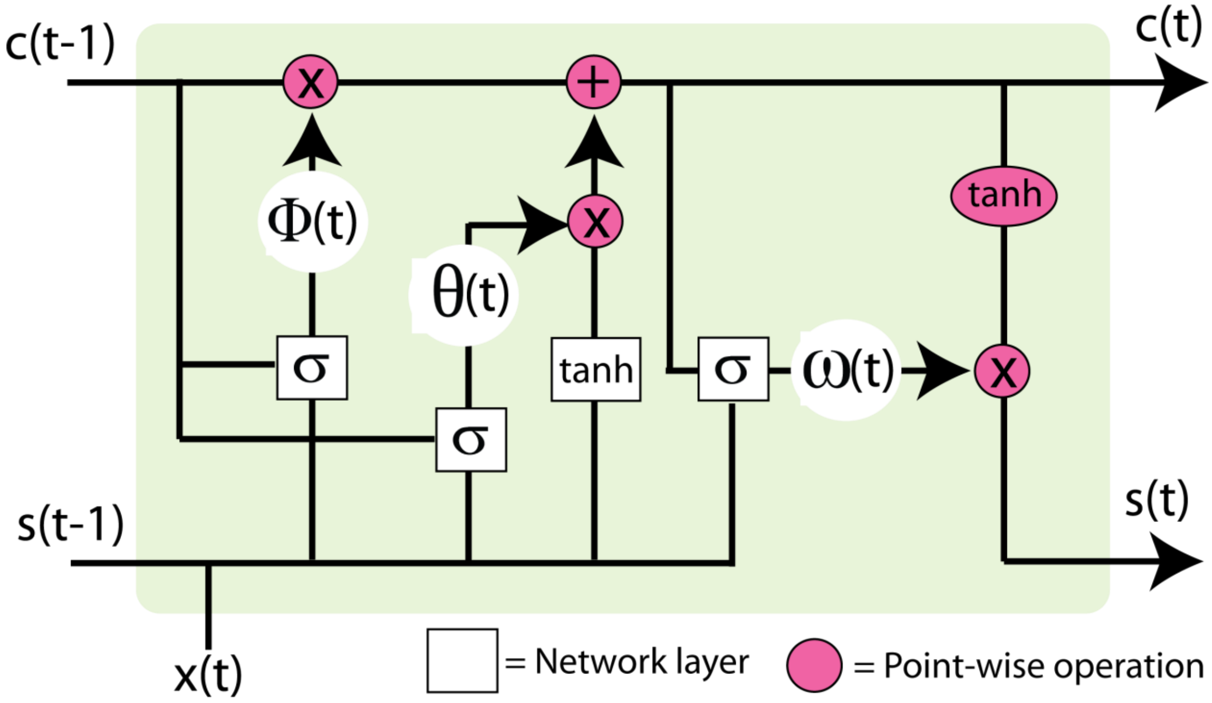


Figure 2-6: A schematic illustration of LSTM architecture (see Equations (13) – (18))

To understand how this architecture effectively prevents the error gradient from vanishing as it passes through more layers, we need to see how the gradient back-propagates from t to $t - 1$ in the cell state. From (17), it is clear that the hidden layer activation in LSTM $s(t)$ is determined by the cell state $c(t)$. Therefore, we just need to prove that the gradient does not necessarily diminish from $c(t)$ to $c(t-1)$. Using an arbitrary loss function $H(Y, O)$ and a chain rule, the BPTT can simply be expressed as:

$$\frac{\partial}{\partial c(t-1)} H(Y, O) = \frac{\partial H(Y, O)}{\partial c(t)} \circledast \emptyset(t) \ldots (19)$$

From (15), $\emptyset(t)$ is a forget gate activation which controls for the rate at which the neural network forgets its past memory. Hence, (19) simply follows from (15) defining how the new memory content at t is constructed: note that there isn't any non-linear activation function involved in generating this new content. In other words, the new memory content is generated from an identity function on the weighted combination of the previous cell state and the current input activation in the hidden layer. As a result, the error gradient does neither exponentially decrease (i.e. the derivative of an identity function is 1) nor explodes (i.e. the forget gate activation, which is basically a vector of sigmoid weights, is always less than 1) even if it passes through a number of previous cell states. The gradient is only linearly modulated by the forget gate activation $\emptyset(t)$. This is how LSTM architecture can preserve the long distance dependency information in its memory if it decides to.

2.4. Quantifying the "degree" in prediction: the information-theoretic framework

Under the view of prediction as a probabilistic phenomenon, constraint can be expressed in the form of the probability distribution. Such probability distribution captures various possibilities with different degrees of expectation which can be compared with the other probability distributions associated with different linguistic contexts in order to illuminate how the processing state of a system changes as a function of prediction. However, we can ask a more fundamental question: Is the constraint useful? In fact, it is not absurd to think that the human language system is flexible to utilize the constraint only if it is informative enough. If the constraint is not very informative, there is really no point to change the processing state. Information theory (Shannon, 1948) offers a way to quantify the amount of information contained in the constraint in the form of a probability distribution, providing an answer to the above question.

One of the key measures in information theory is known as "entropy" which quantifies how much uncertainty is involved in the value of a random variable or the outcome of a random process. The total number of bits (common currency in information theory) is defined by the expected value of the negative logarithm of the probability mass function (PMF):

$$H(Y) = E\big[-\log\big(P(Y)\big)\big] = -\sum_{i=1}^{N} P(y_i)\log P(y_i) \ldots (20)$$

where Y is a random variable with N possible outcomes. The logarithm of a probability distribution is often very useful as it renders the computation additive for independent sources: for example, if the entropy of a fair coin toss is 1 bit, the entropy of m tosses is simply m bits. Due to this effect, the logarithm is commonly adopted to maximize a likelihood or posterior in many statistical optimization algorithms described throughout this thesis. To make the interpretation more straightforward, consider a coin toss. The entropy (uncertainty) is at its maximum if the coin is fair (i.e. the distribution is uniform) because knowing that the coin is fair does not help a system to make a correct prediction at all. However, if the coin is unfair such that one outcome is more probable than the other, knowing the actual probabilities associated with these outcomes clearly improves the prediction (and the entropy becomes lower). Using entropy as a model of human speech comprehension allows researchers to test the hypothesis that the entropy is incrementally tracked throughout the speech such that the prediction only occurs when the constraint is informative (i.e. when the entropy is low). In the context of incremental speech comprehension, the constraint entropy naturally decreases as more words are heard in a sentence because the constraint often becomes more informative with the richer context. This tendency is known as entropy reduction, an important descriptive property of incremental speech comprehension (Hale, 2006).

Entropy describes the degree of uncertainty within a probability distribution, then, cross-entropy measures the expected number of bits that will be needed to predict an upcoming input linguistic unit using an estimated distribution instead of a true distribution. As a result, the cross entropy will always be higher than entropy because using the estimated constraint will always require extra bits than using the true constraint (in the context of incremental speech comprehension, the estimated and the true constraints refer to the prior and the posterior of the belief updating system as illustrated in Figure 2-1). It consists of two terms: the entropy of the true constraint (minimum number of bits required for prediction) and the KL-divergence between the true and estimated constraints (extra bits additionally required for prediction if you are using an estimated distribution):

$$H(Y,O) = H(Y) + D_{KL}(Y||O) = -\sum_{i=1}^{N} P(y_i)\log P(o_i) \ldots (21)$$

where O is the estimated distribution of Y (see (10)). As described above, the cross-entropy is a common error function in neural networks with the softmax activation in the output layer where the softmax output is the estimate of a true distribution. If the true distribution is delta (or a label), then, the cross entropy function becomes equivalent to surprisal.

Computing the entropy of the constraint enables us to quantify how informative it is to predict an upcoming input. This metric could be the basis of deciding whether to utilize the constraint or not. Then, can we quantify the effect of prediction on processing the upcoming input? This is another critical question that could advocate prediction as a core speech processing mechanism in humans. Conceptually, it is not very difficult to formulate a model to address the question: how unexpected is the outcome given the prediction? This can be quantified by any distance function between the prediction O and the outcome Y. In the information theoretic setting, we use the forward KL divergence between these two distributions: $D_{KL}(Y||O)$. If the outcome Y is a label representing the target word being heard, Y always consists of 1 for the target and 0 for all other words that have been considered in prediction O. Then, the effect of prediction on processing the target can be formulated as:

$$D_{KL}(Y||O) = \sum_{i=1}^{N} P(y_i)\log\frac{P(y_i)}{P(o_i)} = \sum_{i=1}^{N}\begin{cases}1 * \log\left(\frac{1}{P(o_i)}\right) & if\ i = j \\ 0 & otherwise\end{cases} = -\log P(o_j) \ldots (22)$$

where j is an index of the target word in the distribution. This simplification is known as "surprisal", reflecting how difficult it is to process the target with respect to the given context (i.e. if the target o_j is strongly predicted such that $P(o_j)$ is high, $-\log P(o_j)$ is consequently low and vice versa). Using the same logic, it is possible to model the belief (prediction) updating process as each word incrementally unfolds in a sentence (see multicycle BBU framework in 2.1). It is merely the KL-divergence between the constraints before and after taking a new input into account. If a new input does not affect the state of belief at all, then, the constraint will not change even after taking the new input into account. However, if it does affect, the degree of update will be quantified under this formulation. From here on, I refer any metrics that represent "how different the target linguistic unit is with respect to the prior constraint" to constraint error (hence, this is not a term to describe the quality of constraint) and surprisal is a particular way to represent the constraint error using KL-divergence.

Referring back to the cross-entropy (21) often used as a loss function in training neural networks (10), if the posterior distribution $P(Y)$ is simply a label indicating a target, the KL-divergence simplifies to (22) and the posterior entropy $H(Y)$ becomes 0 because there is no uncertainty. With a jth response being the target, it is not very difficult to translate (21) to (22). This is why the cross entropy is known as a generalized metric of surprisal and is commonly used as a loss (error) function in many training algorithms.

It has long been claimed that the subjective experience of stimulus intensity is proportional to logarithm of the actual objective intensity (see Appendix 3 for Weber-Fechner's law motivating logarithm as a psychophysical mapping function). In line with this claim, a recent psycholinguistic study revealed that the reading time is logarithmically related to the objective prediction derived from a corpus-based computational model (Smith & Levy, 2013). The surprisal metric has been applied in the field of psycho- and neuro-linguistics and showed that humans are indeed sensitive to the prediction error during language comprehension, providing evidence for prediction as a core mechanism of incremental speech comprehension. See Levy (2008) for theoretical descriptions of information theoretic metrics, Smith & Levy, 2013 for logarithmic approximation of human reading time, Frank et al. (2013, 2015) for application of surprisal for modelling electroencephalography (EEG) data during sentence reading and Willems et al. (2015) for application of surprisal for modelling fMRI data during sentence listening. In this thesis, the information theoretic (logarithmic) models are central to the univariate analysis of neural response amplitude consistent with the abundant applications of the surprisal metric in the psycho- and neuro-linguistic literatures (Roark, Bachrach, Cardenas & Pallier, 2009; Frank & Bod, 2011; Fossum & Levy, 2012; Smith & Levy, 2013; Monsalve, Frank & Vigliocco, 2012; Frank et al., 2013, 2015; Willems et al., 2015).

2.5. Constraints modelling

2.5.1 Modelling a constraint on syntax

One of the most intriguing aspects of language in human cognition is that a word contains multiple levels of linguistic information that allows comprehenders to update their structural interpretation at the message level. Psycholinguistic theories explaining how syntactic knowledge can influence the interpretation of an upcoming word are discussed in Chapter 1. For example, nobody interprets "*shot*" in "*Take the shot*" as a past or past-principle form of a verb "*shoot*" given that a determiner "*the*" can never be a specifier of a verb phrase (plus a verb phrase cannot have another verb phrase in its maximal projection unless a complement phrase bridges them at the intermediate projection). This simple example illuminates how knowledge-based grammatical parsing could provide a useful insight into how comprehenders interpret the sentential structure.

Computational models of grammar select one parser and process one or more corpora with it. The output is often in the form of a probability distribution on which the information theoretic metrics can operate (i.e. the parser's interpretation is the one with the highest probability). In this thesis, I used the VALEX lexical database providing a probability distribution over 163 possible subcategorization frames (SCFs; Korhonen et al., 2006) to model the syntactic prediction in humans at the point of a main verb in a sentence. VALEX is a large lexical database providing lexicalized SCF information for 6,397 English verbs created by processing about 15.9 million sentences extracted from 5 different corpora using the "robust accurate statistical parser" (RASP[2] ; Briscoe & Carroll, 2002). The main verb is a central hub of the sentence on which most grammatical analyses are initiated by informing a particular set of syntactic arguments with which it can co-occur (known as SCFs). This information is an essential component of the lexical functional grammar as it directly constrains the grammatical functions associated with a lexical unit (e.g. verb in this case). Using this probabilistic model of SCFs, I investigate how listeners utilize this core syntactic information to constrain the syntactic structure of the verb complement during incremental speech comprehension.

On the other hand, using a behavioural model allows us to manipulate the richness of the context (from a discourse to a single word) to investigate various sources of constraints which could either converge or conflict (see Marslen-Wilson et al., 1993). Due to the combinatorial explosion in language, it is often difficult to construct reliable constraints based on corpora as the context size grows unless one uses a more sophisticated model like RNN/LSTM (2.3 in Chapter 2). Therefore, not only do behavioural models have a distinct advantage that corpus-based models do not, but they also allow researchers to investigate the effect of cumulative constraints in relation to lexical constraints. This addresses some interesting questions like how the lexical constraints based on a single word (e.g. verb) are neurally expressed when a word is heard in a constraining context. To model syntactic constraints on the verb's complement, I ran a behavioural study in which 15 participants heard the full context consisting of a subject noun phrase and a verb (e.g. "*The experienced walker chose* …"), and provided a probable continuation that came to their mind. Then, their responses were coded in terms of the complement structure used and the occurrences of each of the SCFs were counted and normalized by the total frequency across all frames to generate a probability distribution. In this thesis, this probability distribution for each context is used as a quantitative model of the syntactic constraints provided by the full-context consisting of both the subject noun phrase and the verb. It is compared with the VALEX model reflecting the syntactic constraint based on a verb-alone to demonstrate the relative importance of lexically driven constraint during incremental sentence processing.

2.5.2. Modelling constraints on semantics

The ultimate goal of communication is to understand the message that speaker intends to convey. Semantics is a study of meaning in linguistics which constitutes the message in the context and environment that people are communicating. Therefore, it has been a rigorous topic to define semantics as a representational property in the field of cognitive science. Perhaps, the most intuitive and appealing approach is to characterize the semantic representation by features shared among linguistic objects (McRae, De Sa & Seidenberg, 1997; McRae, Cree, Westmacott & De Sa 1999; Devereux et al., 2014). Assuming that the conceptual knowledge of these objects is organized by the features, the representation defines semantics of each object such as "*sofa*", "*cat*" and "*cabbage*" through the knowledge structure consisting of hierarchical categories such as "*furniture*", "*animal*" and "*vegetable*". Statistical characteristics in the features have been proposed as fundamental principles of cognitive models and used to model the conceptual representation in the neural activity during visual object processing (Clarke, Taylor, Devereux, Randall & Tyler., 2013).

Despite its theoretical appeal and wide applications in the field of cognitive and brain science, it has an important downside in application to modelling language processing. Incrementality is one of the key aspects in human language processing which allows flexible interpretation of a linguistic object with respect to its preceding context. For example, the conceptual knowledge that "*lion*" is a predator does not help to process "*The giant crocodile attacked the lion trying to cross a river*". Semantic understanding of this example sentence requires flexible modification on the underlying conceptual knowledge of "*lion*" as prey.

Unlike the way that a sentence is understood in many theories of grammar, speech comprehension in practice can be facilitated by top-down constraints to process the rapidly unfolding inputs as efficiently as possible. As described in the beginning of this chapter, behavioural studies have shown that the degree to which the upcoming input is predicted entirely depends on how constraining the context is. For example, in a highly constraining context like "*The day was breezy so the boy went outside to fly a …*" (DeLong et al., 2005), the prediction is likely to propagate to the perceptual level (lexical-phonological) compared to less constraining or under-developed context like "*Flying a …*". Nevertheless, even a poor context can still constrain a few semantic features that can co-occur. Such semantic constraints are especially useful in updating the message via interaction with the bottom-up input during incremental speech comprehension in a predictive framework (see 2.3.2).

This motivates the distributional semantic modelling (DSM) approach which captures the statistical relation among words (or linguistic units) with respect to the co-occurring context, under the fundamental assumption that semantically similar words appear in similar contexts (distributional hypothesis; Harris, 1954). In this thesis, I refer to any models that are built upon the distributional hypothesis as DSM. DSM is one of the most popular semantic modelling approaches in computational linguistics as it enables the semantic contents to be induced from the statistics of large-scale text corpora. In this distributional perspective, any words that are conceptually opposite such as "*forget*" and "*remember*" can be very similar because they are occurring in similar contexts. Unlike the feature-based conceptual semantic models, this approach could characterize different aspects of meaning constrained by varying linguistic positions (e.g. semantics of "*lion*" as an object of a verb). In this section, I describe different approaches to compressing the constraint at a lexical level to a semantic level in the DSM framework (these approaches may not be a standard DSM, but they are still in the DSM framework as they are built upon the distributional hypothesis).

2.5.2(a) Modelling constraint in the conceptual hierarchy

As briefly discussed above, typical feature-based semantic models do not capture incrementality: one of the most important aspects in human language processing. Instead, DSM has gained attention from many computational linguists through its appealing traits (Baroni & Lenci, 2010). The primary goal of common DSM approaches is to characterize semantics through the usage of different words in the linguistic environment by comparing a pair of distributions associated with different words. The model developed in this section aims to take an advantage from both sides (i.e. conceptual semantic modelling vs. DSM), capturing the distributional properties of different words (i.e. verbs) by defining the distributions through a set of clearly interpretable semantic concepts. This model is a DSM variant because the co-occurrence data (between a verb and nouns in its complement) was taken as an input (the algorithm projects such co-occurrence data to the conceptual hierarchy and finds the optimal cut at which the representational cost of a distribution (or a verb's semantic constraint on its complement) is at its minimum). Note that the co-occurrence data was obtained from the VALEX database which organised the frequency each verb with possible co-occurring nouns in different subcategorization frames. This section aims to describe every step involved in generating this model, providing a verb's semantic constraint on its complement with a (optimized) set of semantic concepts (see Figure 2-7).

Similar to the SCF constraint in syntax, the semantic constraint can be represented as a probability distribution over a conceptual hierarchy (McCarthy, 2001). For example, the verb "*eat*" would constrain its complement semantics to be about food, having a distribution over different types of foods in conceptual space (Hare et al., 2003, 2004). I borrowed such conceptual space organized into a large hierarchy of concepts from WordNet (Miller, 1995). It is a large database in which conceptual space is defined with each node in the hierarchy, called synset (i.e. node = synset), being linked to the other nodes by means of a small number of conceptual-semantic relations. Although this may sound like an up-side-down tree with every node in the leaves eventually converging to the entity node in the root, it is a more complicated directed acyclic graph (DAG) in reality due to every node in the leaves having one or more connections to the upper level of the hierarchy. Now, the problem reduces to projecting the lexical (word-level) constraint to this WordNet hierarchy (Again, the lexical constraint was given by the VALEX database which provides the frequency of the possible nouns in a particular SCF frame with a preceding verb).

The procedures involved in obtaining a model of constraint represented by an optimized set of synsets are described in Appendix 5 in detail and a simplistic overview is shown in Figure 2-7. Compared to more typical DSM approaches described below in 2.5.2(b), this WordNet approach provides much clearer interpretation of each dimension (or feature) of the constraint. Since the optimization scheme was applied to each of the verb independently, the hierarchical level of representation naturally varies depending on how informative a verb is in constraining its argument such that the verb with a more informative constraint is represented with more specific synsets at the particular region in the WordNet space that the verb prefers (e.g. “suffer” prefers the regions associated with disease, illness or disorder). This optimization scheme, independently applied to each verb, is also an important advantage of this model over the other typical DSMs (However, the downside of this particular aspect when analyzing the data in the RSA framework is discussed in Section 3.6.3 in Chapter 3). The output constraint defined across 15 synsets (most commonly represented synsets across 50 different verbs for comparison) is shown in Figure 2-8.

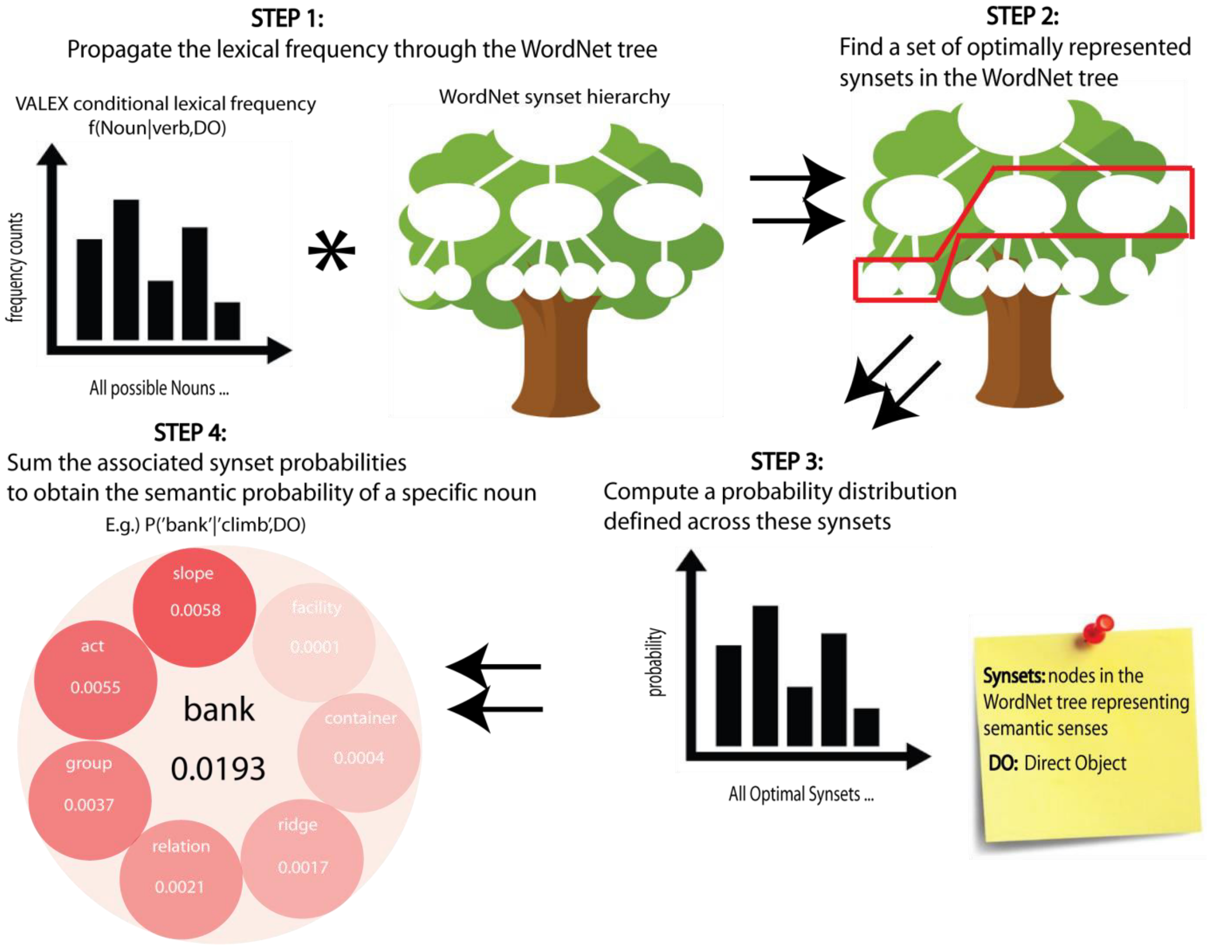

STEP 1:
Propagate the lexical frequency through the WordNet tree
VALEX conditional lexical frequency
f(Noun|verb,DO)
frequency counts
All possible Nouns ...
*
WordNet synset hierarchy
STEP 2:
Find a set of optimally represented synsets in the WordNet tree
STEP 3:
Compute a probability distribution defined across these synsets
probability
All Optimal Synsets ...
Synsets: nodes in the WordNet tree representing semantic senses
DO: Direct Object
STEP 4:
Sum the associated synset probabilities to obtain the semantic probability of a specific noun
E.g.) P('bank'|'climb',DO)
slope
0.0058
facility
0.0001
act
0.0055
bank
0.0193
container
0.0004
group
0.0037
ridge
0.0017
relation
0.0021

Figure 2-7: A schematic overview of different steps involved in generating a semantic constraint model in a hierarchical conceptual space. Note that STEP 3 in this figure already provides the constraint and STEP 4 is only needed when the semantic probability of a specific word from the constraint is requested (e.g. surprisal). Also, note that the WordNet hierarchy is depicted as a tree only for an illustration purpose (it is more complicated DAG in practice).

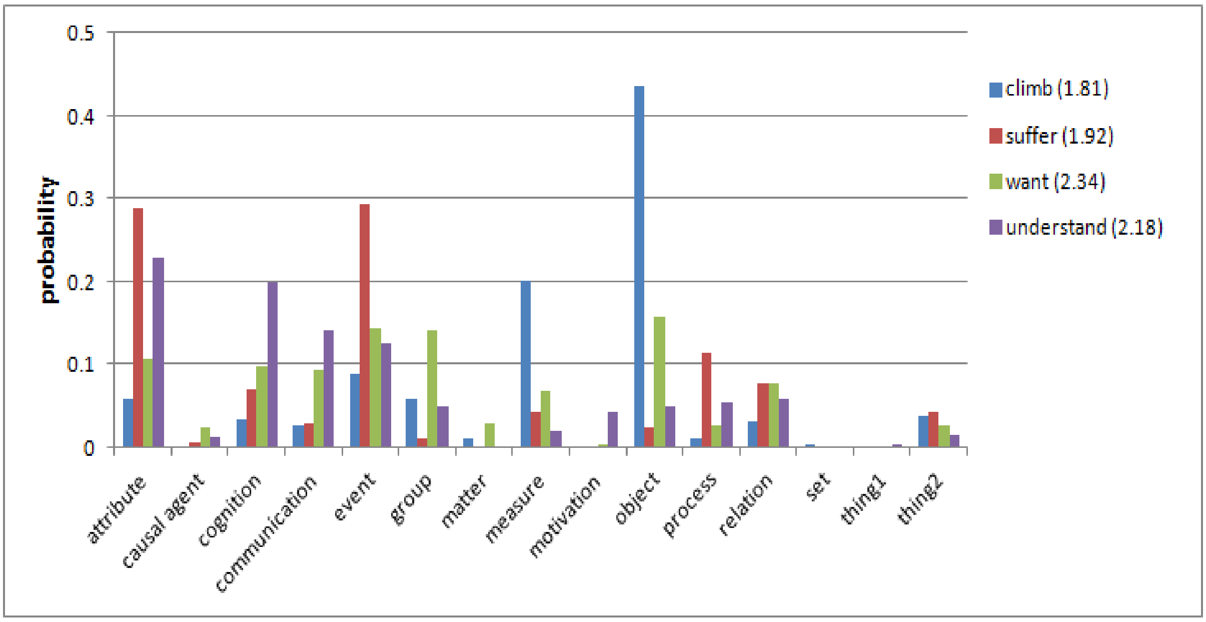


Figure 2-8: Illustration of the semantic constraints defined by the mean optimal cut. The value inside the bracket of each verb in the legend represents the entropy of the distribution. As expected, the constraining verbs like "climb" and "suffer" have low entropy compared to less constraining verbs like "want" and "understand".

2.5.2(b) Latent semantic modelling

Other than the co-occurrence based semantic model defined in the conceptual hierarchical (WordNet) space (see 2.5.2(a)), more typical distributional semantic models were also constructed to capture the semantic content and constraint activated by a word. Despite having dimensions that are not as clearly interpretable as the model defined in the WordNet hierarchy, such semantic models have often been used to capture similarity among different words in terms of their distributional properties (hence, interpreting the distribution as a whole and characterizing semantic similarity through comparing a pair of distributions associated with different words have been the main research topics in such models). In this section, I review different branches of DSM that are commonly employed in the literature and used in this study.

Similar to the conceptual semantic modelling which projects the lexical constraint to the pre-defined conceptual semantic space, latent modelling projects the lexical constraint to the latent space, consisting of a set of dimensions each of which reflects a cluster of words occurring in similar contexts. The total number of dimensions is always smaller than the total number of contexts in the corpus so that the content of each word is efficiently captured as a distribution of a manageable size. The most straightforward approach is to use one of the dimensionality reduction techniques which project the data to a smaller set of orthogonal dimensions while preserving as much variance in the original data space as possible. This type of approach first organizes the corpus data into a matrix of co-occurrence scores whose covariance can, then, be input to a dimensionality reduction technique.

For example, Baroni and Lenci (2010) organized their co-occurrence data to the weighted tuple structure $t(w)$ consisting of a set of two content words w_i and w_j connected by a co-occurrence link l: $t(w) = \{[w_i, l, w_j], v_t\}$. v_t is the co-occurrence score associated with the tuple structure. They used local mutual information (LMI; see (25) below) value as the co-occurrence score reflecting the raw co-occurrence frequency O_{ilj} weighted by point-wise mutual information (PMI) $\log\frac{O_{ilj}}{E_{ilj}}$ where E_{ilj} is the expected count of the same tuple under independence. It is mutual information specific to the tuple $[w_i, l, w_j]$ reflecting the strength of association among the three components after controlling for their individual frequency. They labelled and matricized all tuples into $|w1|$ rows and $|L| * |w2|$ columns where $w_i \in w1$, $l \in L$ and $w_j \in w2$ and used singular value decomposition (SVD) to compress the sparsely distributed data across $Lw2$ column space (see [3] below for a set of co-occurrence links L). SVD finds the orthogonal subspace spanned by the $Lw2$ basis vectors in $R^{|w1|}$. Then, it is possible to single out the basis vectors which do not contribute much to explaining the variance and remove them based on their associated singular values. This is especially the case because the projection does not lose any variance as long as the orthogonal subspace of $R^{|w1|}$ is spanned by $|w1|$ number of basis vectors. The selected set of m basis vectors in the right singular matrix are, then, used to project the original data to R^m orthogonal subspace, generating $|w1|$ by m reduced tensor matrix. This output matrix from Baroni and Lenci (2010) was used as a model of co-occurrence semantics in this thesis.

Since concatenating L onto $w2$ renders the model to reflect the semantic content of a word generalized across possible co-occurrence links (see [3]), selecting a subset of the co-

[3]: A number of syntactic relations in L

Below shows a number of syntactic relations (underlined) organized into a tuple. The example phrase or sentence is given at the end in *Italics*.

occurrence data with a particular link makes the model more specific to the syntactic position in a sentence. This was particularly useful in my analysis in which the epoch of interest was aligned to the main verb of each sentence. Using their tensor data, I trained my own model of semantic constraint specifically in a direct object frame under the topic modelling framework described below.

Topic modelling in a Bayesian framework

Another approach to latent semantic modelling develops a generative probabilistic model which assigns a word to different latent dimensions in a way that maximizes the likelihood or posterior of the model. This probabilistic framework is intuitively appealing because it fits the model to data directly as in regression problem. Typical probabilistic models of language consist of a mixture of context (i.e. predicate) and word (i.e. argument) components across latent variables known as ‘topics’ such that:

$$P(w|c) = \sum_{z} P(w|z)P(z|c) \ldots (23)$$

where w, c and z represent the word, context and topic respectively. The above equation (23) is based on a conditional independence assumption that w and c are conditionally independent given z which is common in models with latent variables such as Hidden Markov Models (HMMs). This framework was used to model the semantic constraints of a verb (c) on its complement noun (w). This modelling approach can be understood in the light of latent Dirichlet allocation (LDA) in a Bayesian framework.

The training samples were obtained from Baroni’s distributional memory (DM) tensor data (Baroni & Lenci, 2010) which organized the co-occurrence into a tuple with a number of different syntactic relations. In order to prevent any confounding effects due to the difference in the subcategorization frame preference between different verbs, I constrained the frame to be a direct object. An important limitation of co-occurrence data with raw frequencies is that

- Noun modifier relation: [good, nmod, teacher] = “*good teacher*”
- Subject argument relation: [soldier, verb, book] = “*The soldier is reading a book*”
- Direct object relation: [book, obj, read] = “*The soldier is reading a book*”
- Indirect object relation: [woman, iobj, give] = “*The soldier gave the woman a letter*”
- Noun coordination relation: [dog, coord, cat] = “*A dog and a cat*”
- Transitive subject relation: [soldier, sbj_tr, read] = “*The soldier is reading a book*”
- Intransitive subject relation: [teacher, sbj_intr, sing] = “*The teacher is singing*”

the two co-occurring words can be strongly associated mainly because they are frequent words and not necessarily because they are semantically related: for example, consider "*love the picture*" and "*cherish the picture*". The raw frequency may show that "*love*" is more strongly related to "*picture*" than "*cherish*" mainly because "*love*" is more frequent than "*cherish*". This is the reason behind choosing the cosine distance as a measure of dissimilarity instead of Euclidean because the cosine distance is based on the angular difference between the distributional vectors, not based on their magnitude difference reflecting the raw frequency (see Baroni & Lenci, 2010). Similarly, mutual information normalizes the raw co-occurrence frequency by the raw frequency of each of the co-occurring words under independence and use this score to weight the raw co-occurrence frequency:

$$I(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \ldots (24)$$

In this way, it prevents the strength of association between two words from being contaminated by their respective raw frequency. The LMI score used in the DM tensor data reflects the mutual information specific to a particular position in the vectors:

$$LMI(x,y) = p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \ldots (25)$$

which is a useful variant to apply to the co-occurrence data of raw frequencies. The observation vector (training samples) was created for every frequency score of the LMI values which were normalized and rounded for this purpose.

Posterior estimation of a multinomial model parameter

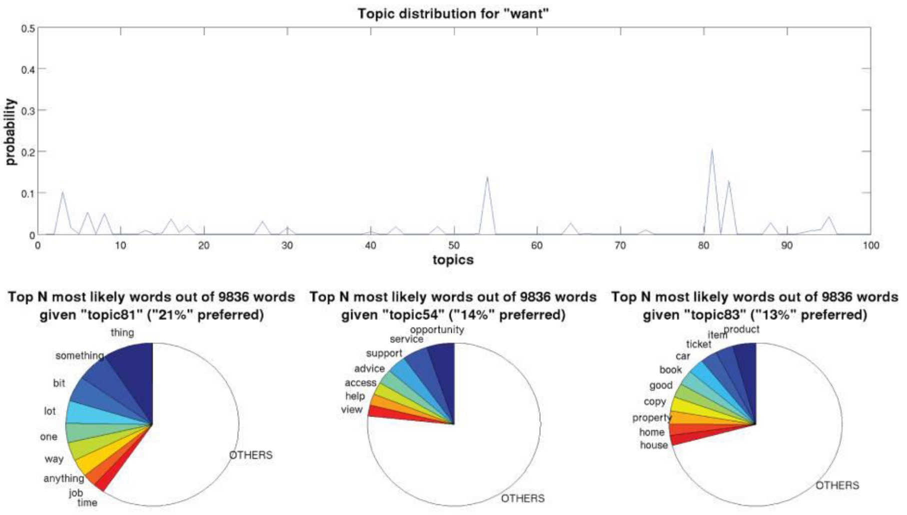
The topic modelling approach is theoretical appealing and conceptually intuitive way of modelling the distributional semantics as the model is trained to fit the co-occurrence data as much as possible. It differs from the variant of latent semantic analysis approach in Baroni & Lenci (2010) where the co-occurrence data was compressed in a way that maximally preserves the original variance in the lexical space using SVD. A simpler analogy between these two approaches is regression vs. principle component analysis (PCA) as different methods of analysing and understanding the data. Just as SVD is generalized variant of PCA in which the singular values of the data matrix are simply related to the eigenvalues of the covariance matrix via the square function, the topic modelling approach finds the model parameters through estimating the posterior as below (similar to the regression where the parameters are typically estimated to maximize the likelihood function).

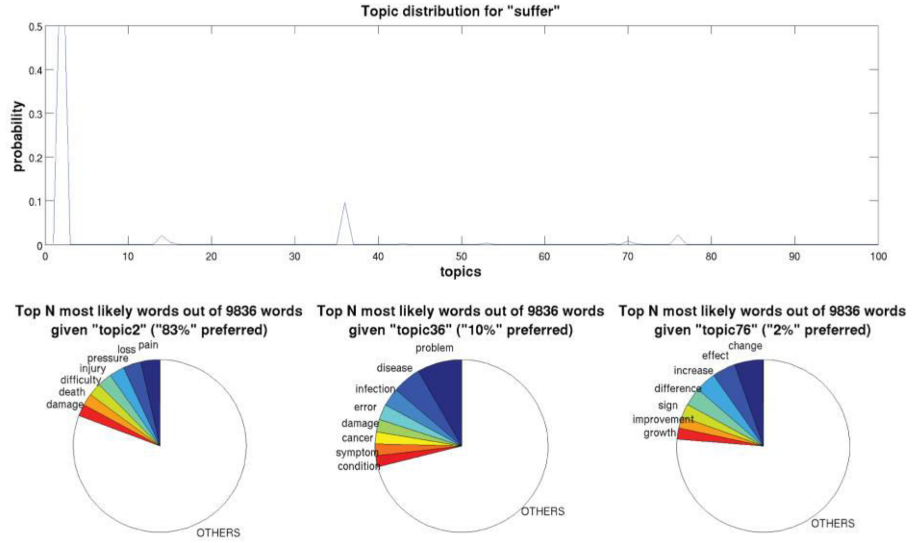
A typical Bayesian approach to probabilistic modelling of a parameter θ based on a given data X is expressed as:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} \dots (26)$$

This formulation is very useful to model a process of learning (or updating beliefs) through empirical observations of data X. Note that the data term X here represents the samples drawn from a discrete vocabulary space in a corpus. Given this multinomial random variable X, the likelihood $P(X|\theta)$ follows a multinomial distribution parameterized by θ ($X{\sim}Multinomial(\theta)$). The parameter θ is typically learnt through the observations X. Given the multinomial likelihood, setting a conjugate Dirichlet prior renders a posterior to follow the Dirichlet distribution. To constrain the posterior to be Dirichlet, the parameter θ can, in turn, be parametrized by a hyper-parameter α ($\theta{\sim}Dirichlet(\alpha)$). Taking the advantage of using a conjugate prior, it is possible to marginalize the parameter θ and express the Dirichlet posterior in terms of known variable X and the hyper-parameter α. In practice, α takes a value between 0 and 1, determined by the model's prior knowledge about θ: α will be near 1 if it is confident about θ (which will not affect the likelihood in any sense). It practically works as a smoothing parameter on the distribution such that α near 1 leads to a set of all contexts being made up of more topics if θ is a parameter of $P(z|c)$ in (23) or a set of all topics being made up of more words if θ is a parameter of $P(w|z)$ in (23). The statistical model which uses the conjugate Dirichlet prior to estimate the Dirichlet posterior to explain the observations by the unobserved variable(s) is known as latent Dirichlet allocation (LDA).

For model training, I used the collapsed Gibbs sampler approach described in Griffiths (2002; see also, Griffiths & Steyvers (2004); Wallach (2002)). The initial parameter settings were based on O'Seaghdha and Korhonen (2014). Further, a fixed-point iteration approach (Minka, 2000) was used to update the hyper-parameter α in a way to maximize the log-evidence as in O'Seaghdha and Korhonen (2014). To understand the training procedures in detail, see Appendix 4 which clearly describes the mathematical derivations of the training algorithm, showing how each parameter is estimated.


Topic distribution for "want"
probability
0.5
0.4
0.3
0.2
0.1
0
0
10
20
30
40
50
60
70
80
90
100
topics
Top N most likely words out of 9836 words given "topic81" ("21%" preferred)
thing
something
bit
lot
one
way
anything
job
time
OTHERS
Top N most likely words out of 9836 words given "topic54" ("14%" preferred)
opportunity
service
support
advice
access
help
view
OTHERS
Top N most likely words out of 9836 words given "topic83" ("13%" preferred)
item
product
ticket
car
book
good
copy
property
home
house
OTHERS


Topic distribution for "suffer"
probability
0.5
0.4
0.3
0.2
0.1
0
0
10
20
30
40
50
60
70
80
90
100
topics
Top N most likely words out of 9836 words given "topic2" ("83%" preferred)
loss
pain
pressure
injury
difficulty
death
damage
OTHERS
Top N most likely words out of 9836 words given "topic36" ("10%" preferred)
problem
disease
infection
error
damage
cancer
symptom
condition
OTHERS
Top N most likely words out of 9836 words given "topic76" ("2%" preferred)
change
effect
increase
difference
sign
improvement
growth
OTHERS

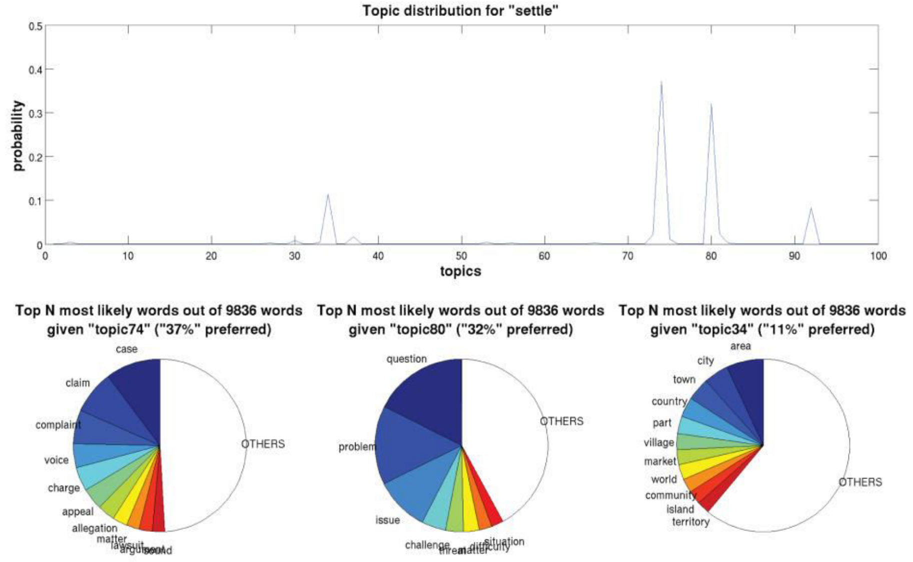

Topic distribution for "settle"
probability
0.5
0.4
0.3
0.2
0.1
0
0
10
20
30
40
50
60
70
80
90
100
topics
Top N most likely words out of 9836 words given "topic74" ("37%" preferred)
case
claim
complaint
voice
charge
appeal
allegation
matter
lawsuit
argument
sound
OTHERS
Top N most likely words out of 9836 words given "topic80" ("32%" preferred)
question
problem
issue
challenge
threat
matter
difficulty
situation
OTHERS
Top N most likely words out of 9836 words given "topic34" ("11%" preferred)
area
city
town
country
part
village
market
world
community
island
territory
OTHERS

Figure 2-9: Visual illustration of the semantic constraints represented by 100 topics. Each set of a distribution and three pie charts shows the topic preferences of a verb in the stimuli (the DT distribution in the top panel) and the object nouns that are highly preferred by one of the top three preferred topics by the given verb (the pie charts in the bottom panel from left to right). The top N words in each pie chart are the object nouns that have at least 0.02 probability of occurring in the given topic.

Chapter 3: Decoding the real-time neurobiological properties of incremental speech comprehension

Understanding spoken language involves a complex set of processes that transform the auditory input into a meaningful interpretation. When listening to spoken language, the ultimate goal is not in acoustic-phonetic detail, but in the speaker's intended meaning. This effortless transition occurs on millisecond timescales, with remarkable speed and accuracy and without any awareness of the complex computations on which it depends. How is this achieved? What are the processes and representations that support the transition from sound to meaning and what are the neurobiological systems in which they are instantiated? Research to date provides a broad outline of the neurobiological language system and of the variables involved in language comprehension (see Chapter 1.2), but surprisingly little is known about the specific spatio-temporal patterning and the neurocomputational properties of the wide range of incremental processing operations that underpin the dynamic transitions from the speech input to the meaningful interpretation of an utterance.

My research combines real-time neuroimaging measurements obtained from EMEG with recent developments in multivariate statistics and computational linguistics to probe directly the dynamic patterns of time-sensitive neural activity that are elicited by spoken words, the constraints they generate on upcoming words, and the incremental processes that combine them into syntactically and semantically coherent utterance interpretation. I used computational linguistic analyses of language corpora and behavioural data to build quantifiable models of constraint and of surprisal, where the latter reflect the processing demands of integrating the upcoming word given the properties of the prior constraints. Based on these cognitive models, I characterized the pattern of neural activity involved in various computations that support dynamic processes of incremental interpretation using representational similarity analysis (RSA; Kriegeskorte et al., 2008; Kriegeskorte & Kievit, 2013). The real-time electrophysiological activity was recorded by EEG + MEG (EMEG) and was compared with the similarity pattern of the models to reveal how different information types are encoded in different brain areas over time during spoken sentence comprehension.

In a previous EMEG study, the spatio-temporal dynamics of a word recognition process were characterized using RSA to test quantifiable cognitive models of key analyses including lexical-semantic competition and semantic feature integration (Kocagoncu et al., 2017). They identified the cortical regions supporting the early phonological and semantic competition between cohort candidates as the word is heard, and the dynamic process of converging to a single candidate and its unique semantic representation as the recognition point approaches. In a subsequent study, the authors investigated how the semantic constraints generated by a noun modifier (“yellow”) interact with the processing of the following noun (“banana”) in a two-word phrase (“yellow banana”), using the cognitive models in a RSA framework (Klimovich-Gray et al., 2019). Combining together, these studies illuminate the temporal dynamics of activating the lexical contents and its interaction with the constraints given by the preceding modifier.

Following on from these studies, my research focuses on the constraints of various linguistic properties during spoken sentence comprehension. The major challenge, as for the word-level studies above, is to develop quantifiable measures of the relevant properties of the sentential processing environment. In this thesis, I investigate the syntactic and semantic aspects of the constraints because they are fundamental cognitive properties of the constraints to guide the sentence-level understanding. Through Chapter 3 and 4, I address how utterances are incrementally combined into a meaningful interpretation of the incoming utterance and how this interpretation modulates the processing of subsequent words in the utterance.

Using these methods, I aim to address the long-standing but unresolved issue in studies of sentence- and discourse-level language comprehension of the relationship between the role of syntactic computations and constraints and the role of semantic and pragmatic knowledge in the interpretation of a spoken utterance (Altmann & Steedman, 1988; Altmann & Mirkovic, 2009; Tyler & Marslen-Wilson, 1977; Marslen-Wilson & Tyler, 1980; Marslen-Wilson et al., 1993). Chapter 1 described two contrasting psycholinguistic accounts of language comprehension. The rule-based accounts (e.g. syntax-first) have argued for the initial use of syntactic knowledge to construct a structure based on the grammatical category information of an input and semantic information is, then, processed under the constructed structure (Frazier & Fodor, 1978; Frazier, 1987). In contrast, the prediction-based accounts (e.g. constraint-satisfaction) claimed that the listeners actively constrain the upcoming continuation using a variety of linguistic cues interactively, given by the preceding context (Marslen-Wilson & Tyler, 1977; Trueswell et al., 1994; Altmann & Mirkovic, 2009). This conflict has never been fully resolved largely because the available experimental methodologies were limited and not able to identify the underlying neural systems whose response patterns characterize the temporal profile of these different types of processes and their potential interaction to drive language comprehension.

With a novel combination of brain recordings with high temporal resolution in a millisecond scale, computational modelling of linguistic properties and multivariate pattern analysis to characterize the linguistic information encoded in the brain activity, this thesis separates out syntactic from semantic constraints, as they evolve over a spoken utterance, and explores the pattern of model-fit across different brain regions. Importantly, the cognitive models that test for effects of syntactic and semantic constraints and their integration into the developing sentence are probabilistic and experiential in nature, reflecting the natural linguistic experience in the real world and providing the quantifiable data from which the rich multivariate pattern can be computed. This approach avoids the limitation of relying on categorical distinctions between stimuli which fails to capture the multifaceted richness of linguistic representations and the probabilistic nature of language and illuminates how linguistic constraints develop over time through resolution and integration during natural speech comprehension.

3.1. Overview

Comparing the effects of multi-level constraints generated by the full context and the verb-alone models enables us to determine how far the effects of contextual constraints are genuinely cumulative. If constraints cumulatively develop as syntactic and semantic information in the context is incrementally interpreted over time (Willems et al., 2015; Altmann & Steedman, 1988; Marslen-Wilson & Tyler, 1980), information associated with the initial subject NP (SNP) will be rapidly projected onto the upcoming speech, so that expectations about the likely properties of the complement noun following the verb will depend on the entire preceding subject NP + verb context, and on the event structure it implies. If the subject NP itself constrains likely complements, the generation of these constraints should be reflected in the MEG signal as the subject NP is being processed. On this view, the incremental integration of new input in sentence processing is not driven only by syntax, nor is it driven by purely lexical syntactic or lexical semantic information associated with individual words in sentences. Instead, semantic and broader conceptual discourse knowledge associated with the entire context has implications for the integration of the subsequent input (Altmann & Mirkovic, 2009; Tyler & Marslen-Wilson, 1977; Marslen-Wilson & Tyler, 1987; Marslen-Wilson et al., 1993; Kuperberg & Jaeger, 2016; Kuperberg, 2016; Nieuwland & Van Berkum, 2006; Matusalem, Kutas, Urbach et al., 2012).

Further, these constraint models were tested in conjunction with the models of constraint error in the relevant epochs (see Figure 3-3) in order to clarify the extent to which different types of constraints influence the processing of the complement. If the effects of the constraints show initial activation of the information that predicts the following complement, the constraint error reflects utilizing such information to facilitate the processing of the complement.

Using RSA on source-localized EMEG data enables us to compare the similarity structure of our theoretically relevant models with the similarity structure of observed patterns of brain activity and can reveal distinct information encoded in different brain areas over time (see 3.2. below). I tested for the timing of the model fit generated for these models across different voxels and time within the fronto-temporo-parietal language network (Binder et al., 2009; Price, 2010, 2012). Based on the previous results (Tyler & Marslen-Wilson, 2008; Tyler et al., 2013), verb-alone syntactic constraint is expected to be activated soon after the verb is recognized in the left posterior middle temporal cortex. Similarly, verb-alone semantic constraint as well as the semantic contents of a word generalized across different frames and positions in a sentence are expected to have an effect in the bilateral posterior temporal cortex (Hickok & Poeppel, 2007; Obleser & Kotz, 2009). In contrast, the activation of full-context constraints is expected to involve more complex processes of combining all information associated with individual words in the context. Thus, the full-context constraints are expected to be activated in the regions involved in combinatorial processing, such as left inferior frontal gyrus (LIFG) for syntax and bilateral anterior temporal and inferior frontal areas for semantics around the onset of a verb (Hickok & Poeppel, 2007; Hagoort, 2005, 2013; Jung-Beeman, 2005).

The effects of the prediction error (or mismatch) of an upcoming word given prior constraints have been previously studied by exploring N400. These studies show that the presence and strength of an N400 response is correlated with the degree of mismatch between the actual word and its prior context (Delong et al., 2005; Federmeier et al., 2007; Frank et al., 2015) and is claimed to be localized to fronto-temporal areas centred on LpMTG (Simos et al., 1997; Lau et al., 2008; Khateb, Pegna, Landis et al., 2010; Maess, Mamashli, Obleser et al., 2016). In light of these studies the effect of semantic surprisal was predicted to be located in bilateral temporal regions whereas the degree of syntactic surprisal was predicted to be reflected in LIFG, the region known to be involved in reanalysis due to a less expected continuation (Tyler et al., 2013).

3.2. Decoding the multivariate pattern of neural activity

The ultimate goal of this thesis is to understand the brain activity during spoken sentence comprehension in order to illuminate the processes involved in understanding speech. The brain activity inherently varies over space and time and, thus, understanding the activity involves interpreting the encoded information from multivariate patterns of the activity using the relevant cognitive models, introduced in Chapter 2. Representational similarity analysis (RSA; Kriegeskorte et al., 2008; Kriegeskorte & Kievit, 2013) provides a way to probe the different types of neural computations that support dynamic processes of incremental interpretation. Under the view that representing content (or information) is a primary function of neural activity, the central notion of RSA is to interpret such content in the representational space defined by each neuron over space and time.

3.2.1. Representational dissimilarity matrix (RDM) and distance metrics

Within the dimensions of representational space, the nature of representation is defined by the geometry of individual points reflecting the activity pattern. The coordinates in the space are defined by the activation values of the source vertices and the central notion of RSA is to characterize this "representational geometry" using a distance metric. For every pair of sentences, the representational geometry is compared which generates a distance value between the sentences in a pair which enters to a particular entry in a square symmetric matrix of distance values across pairs of sentences, known as an RDM. Such geometric distance corresponds to the dissimilarity between two patterns and an RDM shows the important distinctions in the sentence stimuli. The utility of an RDM is what made RSA popular in neuroscience: comparing two activity patterns defined in distinct space requires defining a mapping between the vertices in one space to those in another space but comparing two RDMs does not require such mapping as the activity patterns are represented in a form of distance matrices of the same size.

The basic decoding analysis often refers to the linear classification analyses, designed to investigate whether the binary class of a stimulus can accurately be predicted by the neural activity pattern. For this type of analyses to work, the classes must be linearly discriminable to a certain degree; one example of this type is logistic regression in which the classification boundary is defined probabilistically whereas support vector machine (SVM) provides a geometric definition of the boundary. The representational geometry of an item has much richer structure of content beyond the class discriminabilty: the classification of stimuli can be successful for a number of different representational geometries but each geometric pattern exerts its own functional significance in the brain region. Beyond the other decoding analyses, RSA offers a way to compare such rich structure of content existing in the multivariate patterns using RDMs.

Constructing an RDM involves comparing the representational geometry for every pair of item using a particular distance metric. If the representational geometry is defined in a one-dimensional space, calculating the dissimilarity is just as simple as subtracting the activation values. The multivariate version of this absolute distance is known as L1 distance, often used as an objective function in L1 optimization problems. If the representational geometry is defined in the multidimensional space, there are a number of distance metrics having different functional properties. One commonly used example is the correlational distance, defined as 1 – correlation. In this metric, similarity is straightforwardly defined as a degree of relation between two vectors quantified by covariance. Given that covariance is a dot product between the two vectors centred around zero, it is proportional to the magnitude of projection of one vector to the other (e.g. the magnitude of component of X in the direction of Y or vice versa, known as a scalar projection). Using the concept of projection (or projection magnitude to be precise), it is possible to prove that correlation between the two centred vectors is merely a cosine of the angle between the vectors (see Figure 3-1). Therefore, cosine distance, defined as 1 – cosine similarity, is merely a variant of correlation distance which becomes identical given the two mean-corrected vectors.

One of the most commonly used distance metrics between vectors is Euclidean distance. The most standing out property of this distance metric in relation to the aforementioned metrics is that it is sensitive to the length of each vector. Note that the squared Euclidean distance is proportional to cosine distance only if the two vectors are L2-normalized such that $\sum_i {X_i}^2 = \sum_i {Y_i}^2 = 1$:

$$\sum_i (X_i - Y_i)^2 = (X - Y)^T(X - Y) = X^TX - 2X^TY + Y^TY = 2 - 2X^TY$$
$$= 2 - 2cos\angle(X, Y) \ldots (27)$$

This suggests that all of these distance metrics are closely related to each other, depending conditionally on the input vectors. For further details about calculating a noise-normalized distance metric LD_c, see Nili, Wingfield, Walther et al. (2014).

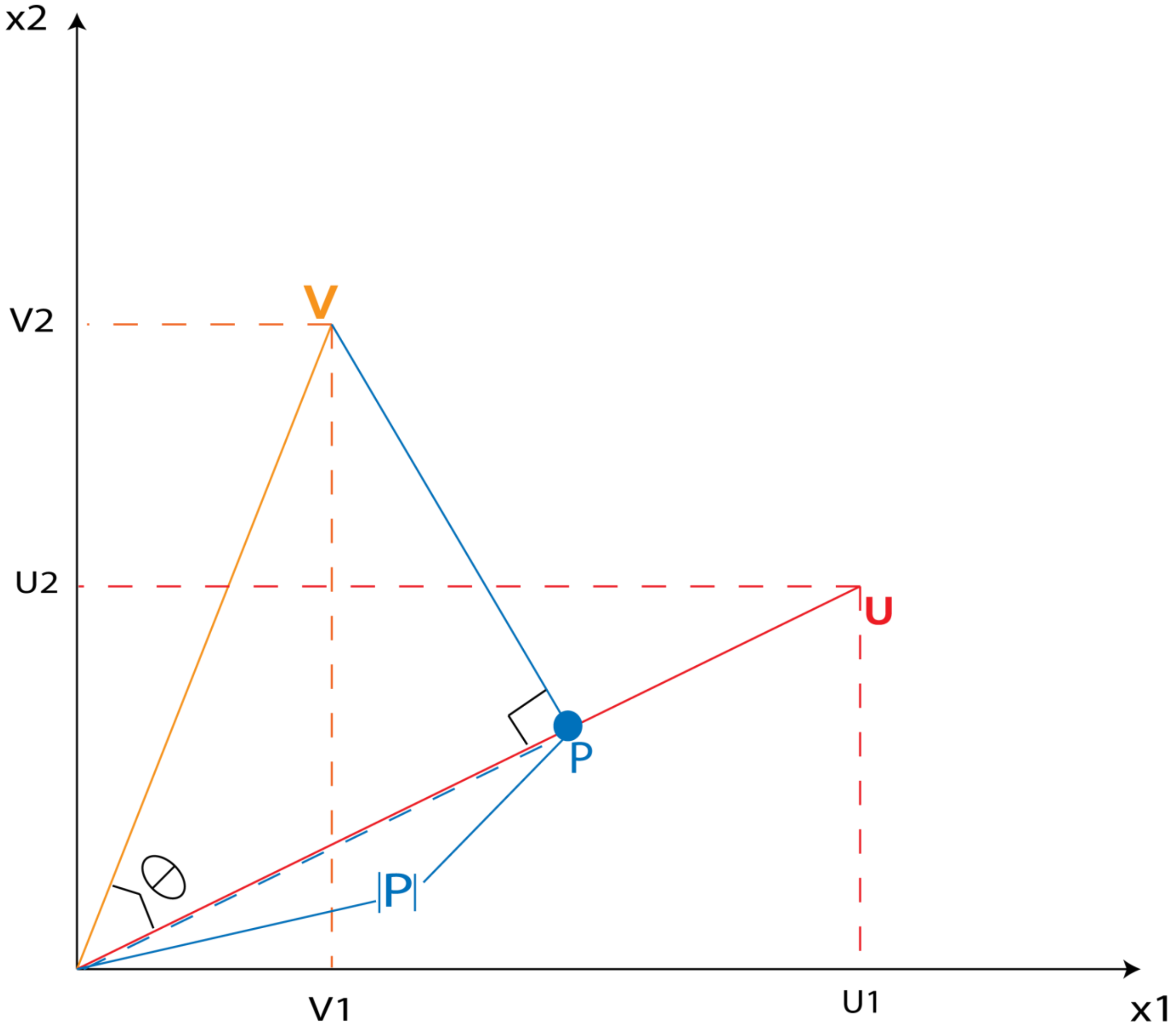


Figure 3-1: A simple illustration of vector projection P, projecting V onto U. There are a number of points to highlight: 1) the magnitude of projection $|P|$ is proportional to the dot product between these vectors such that $|P| = \frac{V \cdot U}{|U|}$ where $\frac{U}{|U|}$ is a unit vector in the direction of U, 2) the vector projection P, interpreted as the component of V in the direction of U, is merely a magnitude of projection applied to the unit vector in the direction of U such that $|P| = \frac{V \cdot U}{|U|} \frac{U}{|U|} = \frac{V \cdot U}{|U|^2} U$, 3) According to the Pythagorean theorem that the cosine of an angle in the right-angled triangle is computed as adjacent divided by hypotenuse, the cosine of θ can be expressed as $cos(\theta) = \frac{|P|}{|V|} = \frac{V \cdot U}{|U||V|}$.

Relating the data to the model RDMs

Choosing a particular distance metric based on its properties is very important when constructing an RDM. For example, neural responses might be consistently higher for one trial than for the other, most likely reflecting the noise. Euclidean distance is prone to this noise returning high dissimilarity value in the RDM. Therefore, I used correlation distance when constructing the data RDM, capturing the dissimilarity based on covariance of neural responses between the two trials regardless of their respective total activation strength. On the other hand, cosine distance was used to construct the model RDMs of semantic constraints to prevent the distance from being affected by the L2 magnitude of the constraint vectors (reflecting how frequently the contexts occur in the corpus). This is why cosine similarity is often employed in computational linguistics: it provides a similarity score based on the association strength between the context and the target word without taking their frequency into account (see Baroni & Lenci, 2010). In case of the model RDMs of syntactic constraints that compare the pair-wise similarities between SCF probability vectors (which only contain 5 specific SCF dimensions), the default Euclidean distance was employed as these vectors were already frequency-normalized. Unlike these constraint RDMs, absolute distance was used to construct the syntactic (SCF) constraint error RDM because the error was quantified by the surprisal metric described in Chapter 2. Lastly, to compute the semantic constraint error RDM, the constraint and the representation of the actual target word were compared using cosine distance as both of them are represented in the same multidimensional space. In order to compare RDMs constructed using different distance metrics, the RDMs were ranked and vectorized to compute Spearman's rank correlation such that the relationship between the RDMs does not have to be linear.

Searchlight analysis over space and time

The source-localized EMEG recordings naturally vary over space and time. In order to investigate various neuro-cognitive processes occurring in parallel over different areas in the brain on millisecond timescale, a data analysis technique known as spatiotemporal searchlight representational dissimilarity analysis (ssRSA; Su, Fonteneau, Marslen-Wilson & Kriegeskorte, 2012) was used. Here, searchlights refer to spheres that span across three dimensional voxel space and one temporal dimension. Each searchlight is defined for each voxel at each time-point, providing a fine-grained spatiotemporal map of where and when in the brain such cognitive processes occur whose representational contents are characterized by different model RDMs. The research topic of this thesis, incremental speech comprehension, naturally involves dynamic processes of constraining, analysing and integrating linguistic units at phonological, lexical, syntactic and semantic levels and expectedly recruits a large distributed neuronal network that includes frontal, temporal and parietal regions (see Price 2010, 2012). By carrying out the searchlight analysis within this large language network, I aimed to elucidate the spatiotemporal dynamics of predictive computations of constraining and integrating an upcoming word at syntactic and semantic levels, using the computational models described below in 3.3. The large language network for the analysis was defined by Harvard-Oxford cortical atlas, a probabilistic atlas created by MNI-registered T1-weighted images of 21 healthy male and 16 healthy female subjects (see Figure 3-2 for surface rendering of this mask).

In order to characterize the spatiotemporal patterns of activation, the data RDMs were constructed from the searchlight spheres with a spatial radius of 10mm (following on from Kocagoncu et al., 2017; Bingjiang et al., in revision) and a temporal radius of 30ms for every 5ms step, so that the data RDMs can capture smoothed patterns of activation over space and time. Then, each of these data RDMs was correlated with the time-constant model RDM, generating a correlation value at each searchlight (see Figure 3-2). This provided a 4-D spatiotemporal map of a model-fit depicting where and when the information encoded in the model is activated in the brain. Once this correlation map was obtained for every subject, I tested if the correlation between the model and the data is consistently above zero across subjects using one-sample t-test at each searchlight. Hence, the map of t-values shows the significant point at which the pattern of neural activity is well characterized by the model of interest consistently across different listeners.

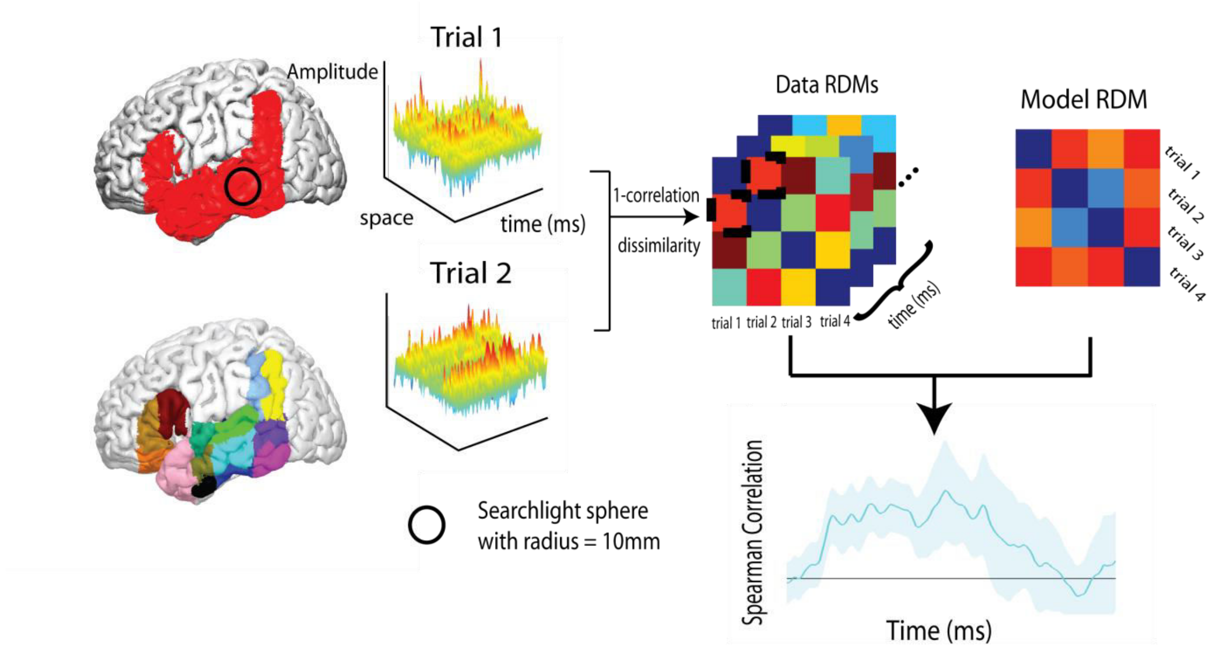


Figure 3-2: A schematic illustration of the searchlight and ROI representational similarity analysis of spatiotemporal source-space EMEG data. The bilateral language mask used in this study is surface-rendered onto the LH brain template in the figure for visualization. Since the source-space EMEG data inherently vary across time and space, I calculated the similarity of the spatio-temporal patterns of brain activities for different trials based on measurements within each searchlight sphere with a spatial radius of 10mm and a temporal radius of 30ms or each ROI with the same temporal radius. I used 1 – Pearson's correlation between pairs of trials as the distance metric to compute a representational dissimilarity matrix (RDM) for each searchlight or ROI, yielding a 4-D map or a time-course of data RDMs. Each data RDM is then correlated with each model RDM (which, in this study, does not change across time) using Spearman's correlation. The figure illustrates this process, yielding a time-course of correlation at a particular spatial location (iterating this process across space will yield the 4-D map of correlation values).

3.2.2. Cluster statistics

In a spatiotemporal map of a neural signal, each time-point and voxel is never really independent of its adjacent time-points and voxels. Not surprisingly, a more elaborate statistical analysis focuses on the cluster(s) of effects, instead of testing the effects at each time-point and voxel independently. A straightforward solution is to set a threshold so that only those searchlights whose t-values are greater than the threshold are used to form clusters via summation. Each cluster level t-value represents the t-values summed across the adjacent searchlights above the threshold. However, such cluster forming threshold (CFT) approach suffers from inconsistent results depending on the threshold value as small variations in the data around the threshold could lead to a large difference in the final output. For example, a spatiotemporally distributed cluster at the threshold p-value of 0.05 could disappear at 0.01 merely because the associated p-value for every time-point is in-between 0.05 and 0.01. Although broader signals are better detected by a low CFT whereas focal signals are better detected by a high CFT (Friston Worsley, Frackowiak et al., 1994), it is difficult to presuppose the nature of spatiotemporal dynamics of incremental computations in the brain.

Hence, I applied the Threshold Free Clustering Enhancement (TFCE) method to take the spatiotemporal clustering of effects into account (Smiths & Nichols, 2009). In comparison with the CFT approach, this TFCE approach aims to optimize the sensitivity to both diffused/extensive and sharp/focal signals by incrementally taking both the cluster-extent and the threshold into account with emphasis parameters manipulating their relative contribution to the output statistical value. From a searchlight map of t-values, the TFCE statistic at a particular searchlight is computed as:

$$TFCE(t) = \int_{h=h_0}^{h_p} e(h)^E h^H dh \ldots (28)$$

where h_0 is the initial threshold, h_p is the maximum threshold, $e(h)$ is the cluster extent at given threshold h and dh is the integration resolution (set to 0.1). I set $h_0 = dh$ and h_p to be the maximum t-value. The emphasis parameters E and H were set to 0.5 and 2 respectively based on the empirical optimization (for E) and the approximation of the log evidence (for H) given that the log evidence is approximately proportional to the square of the statistical threshold h (see Smiths & Nichols, 2009).

As shown in (28), computing the TFCE map from a t-map involves the following procedures. First, calculate how much iteration is required to integrate over the initial and the maximum thresholds (Here, the maximum threshold is set to the maximum t-value in the data and all thresholds are eventually integrated which essentially make this approach “threshold-free”). This can easily be calculated as a number of elements in a vector from the initial threshold to the maximum threshold with the jump of the integration resolution (i.e. $length\left(h_0 \colon dh \colon h_p\right)$). Second, for each iteration step, each element in the vector becomes a threshold and the cluster extent $e(h)$ (i.e. size of the cluster to which each data-point belongs) is computed for every data-point (i.e. searchlight). Note that the cluster extent is necessarily being diminished with increasing threshold and only data-points with high t-value can accumulate the cluster extent for most of the iteration steps (In this sense, TFCE is a "cluster-enhanced", voxel-wise statistics). Lastly, the cluster extent at each iteration step was weighted by each threshold value h in the vector $h_0: dh: h_p$ to emphasize the cluster extent associated with the larger threshold. Then, this weighted cluster-extent was integrated (summed) over all thresholds for each data-point.

The TFCE value can be interpreted as a cluster P-norm $\left[\int_{h=h_0}^{\infty} e(h)^p dh\right]^{1/p}$ which is a generalized Euclidean norm (i.e. a typical measure of vector magnitude with $p = 2$). With a practical discretization of the continuous integral into a finite vector length of CFT increments and with a weighting function w, the cluster P-norm can be expressed as:

$$\left[\sum_{k \in K} w(h_k) e(h_k)^p dh\right]^{1/p} \dots (29)$$

where K is a vector of all CFT increments from dh to h_p. From here, it is clear that dropping the outer-most power and setting $w(h_k) = {h_k}^H, p = E$ gives the TFCE implementation (28):

$$TFCE(t) = \sum_{k \in K} e(h_k)^E {h_k}^H dh \dots (30)$$

summarizing both the t-value at the time-point t and the cluster magnitude to which the time-point t belongs in a form of the weighted summation.

3.2.3. Multiple comparisons correction

The EMEG data naturally varies across space and time and the experimental questions that this thesis addresses involve characterizing the spatiotemporal dynamics of neural activity in different brain regions. As a result, there are multiple independent statistical tests across space and time (note that the spatiotemporal variation within each searchlight was used to capture its representational content). The standard approach for controlling the family wise error rate (FWER; probability of committing a type 1 error) such as Bonferroni correction is too stringent given that the geometric pattern of a regional response for each searchlight is never really independent.

Following on from computing the TFCE map across space and time based on (30), I used the permutation statistics described in Maris and Oostenveld (2007) on the TFCE output. Under the null hypothesis that our model is not correlated with the data, we randomly permuted the sign of correlation values across different subjects and ran one-sample t-test for every searchlight. For each random permutation, this process generated a null 4-D map of t-values which was, then, was converted to a null TFCE map in the same way as above (30). This random permutation process was repeated 1,000 times and the maximum TFCE value across all searchlights was saved for every run. This process generated 1,000 maximum TFCE values under the null hypothesis and the significance of the observed TFCE values were evaluated with respect to this null distribution. This step corrected for the multiple comparisons across space and time without assuming that each sample is independent.

3.3. Models of constraints

In order to decode the multivariate patterns of neural activity involved in understanding speech, I constructed a number of models capturing various linguistic properties of constraints and integration using databases of large-scale corpora. Further, I varied the basis of constraints to investigate the nature of context on which the constraints are conditioned for different linguistic aspects. My primary interest is in what I refer to as "full-context" constraints on upcoming linguistic units, the cumulative constraints generated by the set of words comprising the complete sentential context. The pattern of these constraints were compared with the constraints generated solely by a main verb – called "verb-alone" constraints, designed to capture the lexical nature of constraints activated during spoken sentence comprehension in comparison with the constraints based on the sentential context. This enables us to assess the cumulative effect of the constraints of various linguistic properties generated by the sentence context and to determine how far the constraints activated by the lexical information are expressed in the brain when a main verb is heard in a constraining context.

In line with the other accounts of incremental speech processing, constraints are expected to be computed as the current word is being recognized (Marslen-Wilson, 1975; Marslen-Wilson & Welsh, 1978; Delong et al., 2014). In a natural language environment in which the daily conversation takes place, prior constraints are relatively broad and rich which may favour a specific word as a continuation: "*The day was breezy so the boy went out to fly a …*"

(Delong et al., 2005). However, specific words are rarely strongly predicted (Luke & Christianson, 2016) because such rich context is not always available as in the sentence materials used here: "*The experienced walker chose the* ...". Similarly, a word-level prediction is often very sparse and redundant in computational models which motivates the use of dense clusters through compression for an efficient representation. These computational models are used to construct the model RDMs and tested against the brain data in the RSA framework as described above, primarily focusing on the relative timing with which they appear as the utterance is heard. The effects of constraints on processing the upcoming input and on integrating it into the incremental representation of the prior context are quantified by an information theoretic metric known as surprisal. In summary, the timing and location of the effects captured by these models reveal a picture of when and where the human brain activates and utilizes constraints at both syntactic and semantic levels.

Syntax

Following on from section 2.5.1 in Chapter 2, the output distributions from the VALEX and the pre-test data were used to generate model RDMs of syntactic constraints based on a verb or a full sentential context respectively. For each pair of trials, the Euclidean distance was used to compare the dissimilarity between their syntactic constraints. This distance value was put into a specific entry in the RDM and these model RDMs were tested against the brain data aligned to the onset of the verb in the RSA framework (see Epoch V1 and Epoch V2 in Figure 3-3). The constraint error model was quantified by the surprisal metric (see 2.4) for every sentence stimulus which was, in turn, compared using the absolute distance between every possible pair of stimuli to create the RDMs. This error RDM was tested at the epoch aligned to the onset of the complement to investigate the timing of the update effects in relation to the constraint effects.

Semantics

Note: all semantic constraint models were trained and generated based on the verb and its complement noun co-occurrence specifically in a direct object frame. It is the simplest, yet most frequent, frame in English which enables direct semantic mapping between an agent, a verb and a patient with minimal syntactic intervention. Consequently, any potential confounds due to syntactic variations in the semantic constraint models were removed. Choosing a particular subset of the data allows the trained semantic models to capture the particular semantic aspect (i.e. constraints) specific to the structure of the data. It differs *from the other distributional or feature-based semantic models, capturing the general semantic content of a word which is not specific to the lexical context as in my constraint models.*

Section 2.5.2 in Chapter 2 describes different ways to model the verb-alone semantic constraints based on the verb and object noun co-occurrence data. The WordNet approach provided the verb-alone constraint optimally represented in the hierarchically organized conceptual space in a form of a probability distribution and the distributions associated with different verbs in the trials were pair-wise compared using cosine distance. Similarly, the latent semantic modelling approach provided the verb-alone constraint represented by a set of topics in a form of a posterior distribution of a latent variable conditioned on the verb and the distributions for every pair of verb were compared using cosine distance as in the WordNet approach. The output distance value was entered to a model RDM and compared with the brain data aligned to the onset of a verb (Epoch V2 in Figure 3-3). The WordNet semantic constraint error was constructed by calculating the surprisal value of all synsets (represented at the same level as the constraint) associated with the actual target noun.

Similar to modelling the full-context syntactic constraints, 16 participants were asked to provide the five most probable words that immediately came to their mind after hearing the fragment of the form: subject NP + verb + “the” which signalled a direct object continuation (e.g. “*The experienced walker chose the …*”). The total number of responses for each noun given by these participants were counted and used for modelling the semantic constraints associated with the entire sentential context (any non-noun responses were ignored). Asking the participants for five most likely responses was to improve the reliability of the models by reducing the inter-experiment variability often caused by listeners’ recent experience.

For every unique noun collected from the pre-test, I took the topic distribution of the object noun and weight-averaged the distributions using the pre-test frequency values across the nouns. The goal is to compute:

$$P(topic|context) = \sum_{word} P(topic|word)P(word|context) \dots (31)$$

where

$$P(topic|word) = P(word|topic)\frac{P(topic)}{P(word)} \dots (32)$$

Note that $P(topic|word)$ comes from the topic model and $P(word|context)$ was obtained from the pre-test continuation responses. If all topics are equally probable (this can be checked by computing $P(topic) = \sum_{document} P(topic\ |document)P(document)$ from the same topic model from which the object noun distribution was taken), $P(topic|word)$ essentially comes down to $P(word|topic)$ normalized by $P(word)$; this prevents the blend (31) from being contaminated by the frequency of predicted words. This generates a vector of semantic blend (see Klimovich-Gray et al., 2019), a model of full-context semantic constraint, showing which topics are generally expected by the preceding context based on predicted words from the pre-test. The semantic blend vectors were compared for every pair of the sentential context using the cosine distance, which was entered to the specific entry in an RDM. This RDM was tested against the brain data at the epoch aligned to the onset of the verb (Epoch V1 in Figure 3-3) as well as that aligned to the offset of the context which is same as the onset of the complement noun (Epoch CN1 in Figure 3-3). Only the semantic constraint models were tested both at the onset and the offset given that the target word (complement noun) does not appear straight after the verb. In this way, the analysis using these models at Epoch V1 and Epoch V2 in Figure 3-3 was designed to investigate the earliness with which the constraints are activated whereas the analysis using these models at epoch CN1 was to test how specific such predictive processing is to the target word.

Similarly, the model of semantic constraint error was generated by computing the distance between the topic representations for every predicted noun from the pre-test data and the actual noun in the stimulus sentence (Note that there are multiple distributions associated with many different candidate nouns unlike the verb-alone model). The distance values associated with every predicted noun were, then, weight-averaged using the frequency in the pre-test which reflected how distant the predicted semantics was from the actual semantics of the complement noun. This weight-averaged value was directly entered to the specific entry in an RDM and this RDM was tested against the brain data at the epoch aligned to the onset of the complement noun (Epoch CN2 in Figure 3-3). Note that only 100 direct object sentences were included for this particular model RDM since the topic model was trained specifically to capture semantic constraints in a direct object frame.

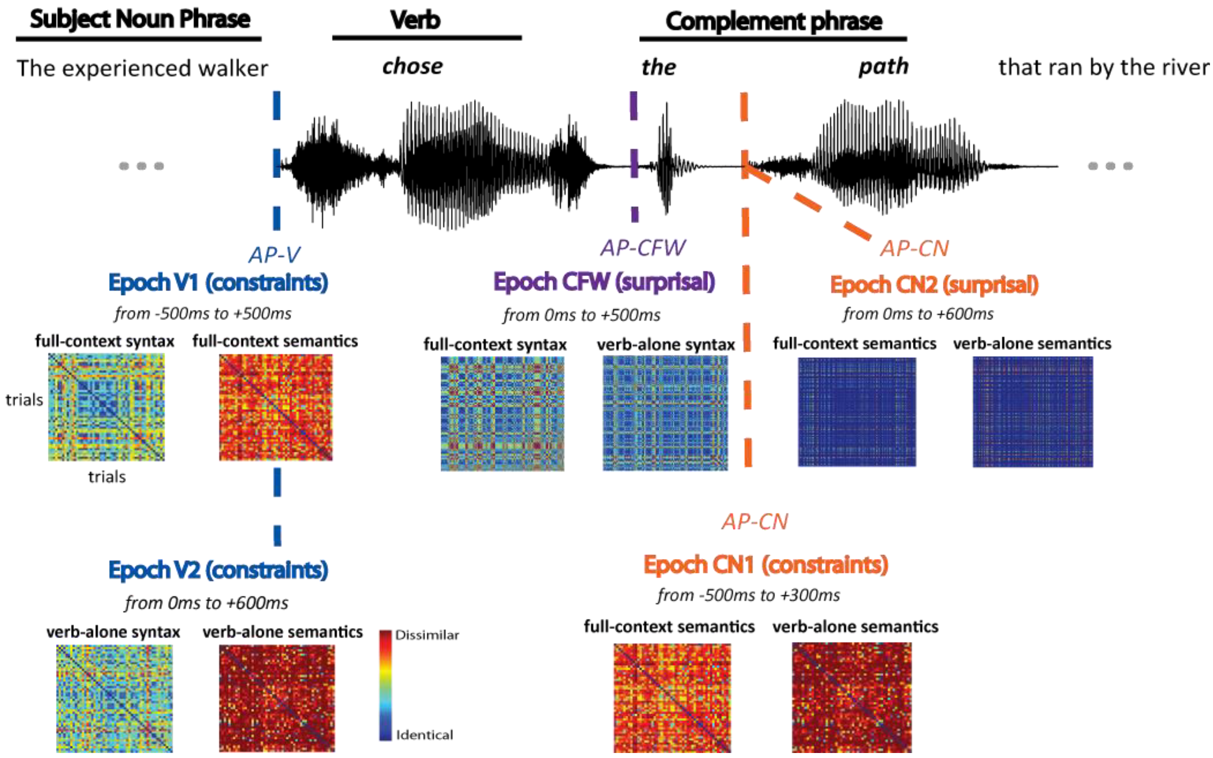


Figure 3-3: Overview of the epochs tested in the experiment in relation to the relevant models of interest associated with each epoch (the models are shown in a form of representational dissimilarity matrices) and to the issues addressed within each epoch. The epochs were each defined relative to an alignment point (AP), with AP-V aligned to the main verb onset ("chose") in blue, AP-CFW aligned to the complement phrase function word onset ("the") in purple and AP-CN aligned to the complement phrase content word onset ("path") in orange. Each AP is marked on the waveform as a vertical broken line. There are five epochs in total (time-window relative to AP given in italics): Epoch V1 and V2 are aligned to AP-V, Epoch CFW to AP-CFW and Epoch CN1 and CN2 to AP-CN.

3.4. Additional analysis: activation of the generalized semantic contents of a constraining word in a sentence

The main topic of this chapter is to decode the underlying properties of the constraints activated while listening to a spoken sentence and their utilization to facilitate processing the upcoming complement. Such constraints depend on a preceding word or a context that can be captured by the statistical regularities in the co-occurrence data in a particular position and a frame. As a result, if the content of a primary source of the constraints, generalized across different positions and frames, is ever activated in the brain, it is expected to be observed in relation to the constraints in an epoch specific to the source. Therefore, I ran an additional analysis capturing the general semantic contents of a subject noun, using the Baroni's DM vectors (Baroni & Lenci, 2010). In conjunction with the topic models trained specifically to capture the semantic constraints, testing this model of the subject noun semantics will highlight the similarities and differences in terms of the timing and regions of activation between the generalized semantic contents and the semantic constraints.

3.5. Results

Using RSA and probabilistic distributional RDMs of syntactic and semantic representations, I probed source-localised EMEG data capturing the real-time electrophysiological activity of the brain to determine the spatiotemporal properties of the cumulative representational context provided by the initial SNP and verb. Full-context models were compared to models restricted to the syntactic and semantic constraints generated by the verb alone. To measure the predictive effects of these representations in the processing of the complement phrase, I used multiple surprisal-based models to examine syntactic and semantic integration effects.

Searchlight Analysis

The cognitive process of constraining and integrating a word at different linguistic levels is very rapid, occurring on millisecond timescale over multiple brain areas. In order to investigate the spatiotemporal dynamics of such predictive computations in the brain, the ssRSA approach was taken to analyze the source-localized EMEG data which vary over space and time. This section does not report any statistics and only presents results to visualize the clusters associated with different models across space and time in the brain. For visualization, each searchlight was used to form a cluster with its adjacent searchlights over space and time only if the p-value (uncorrected) associated with the searchlight was less than 0.01 which was surface-rendered on the brain template. The surface-rendered clusters which are consistent in terms of space and time with the ROI analysis below are highlighted.

Out of all models tested at different epochs described in Figure 3-3, the ones that showed meaningful clusters (despite not being statistically significant) were the full-context semantic constraint, the verb-alone syntactic constraint and the full-context semantic constraint error models. First, the full-context semantic constraint model showed the initial activation of clusters in the right inferior parietal and superior temporal regions around -350ms which transitioned into the right temporal pole (RTP), then to RBA44 around the verb-onset. Although the complement semantics was constrained based predominantly by the RH fronto-temporo-parietal regions based on the subject NP, the LH fronto-temporo-parietal regions (LBA47, LTP and LAG) became involved in constraining the complement semantics around 350ms after the verb-onset (the point after recognizing a verb; see panel (A) in Figure 3-4). Unlike the semantic constraint involving the bilateral language network, constraining syntax primarily recruited the left fronto-temporal regions centred on LBA44 and LMTG (see panel (B) in Figure 3-4a). This cluster first emerged in LMTG and LBA44/45 which peaked around 225ms after the verb-onset. It was transitioned into LSTG/LHG, then to LBA44/45 peaking around 400ms. Lastly, the cluster associated with the semantic constraint error appeared in the posterior temporal lobe peaking around 325ms after the complement noun onset (see Figure 3-4b). This cluster persisted throughout the epoch, possibly reflecting the integration of the object theme carried by the complement noun.

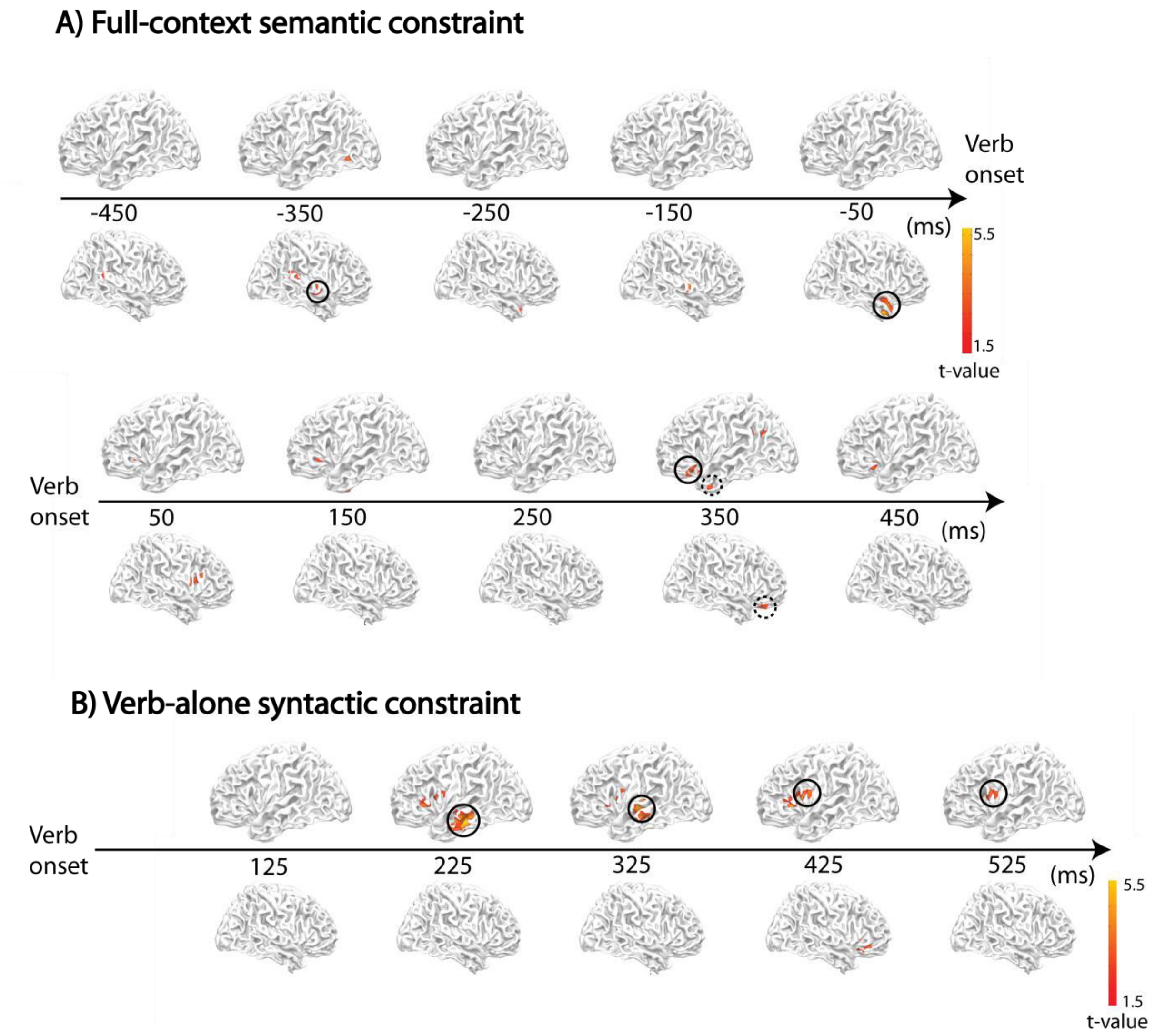


Figure 3-4a: the Searchlight clusters of a full-context semantic constraint (A) and a verb-alone syntactic constraint (B). Any clusters inside the bold circles are consistently observed in the ROI analysis. Similarly, the dotted circles are used to highlight the regions which are marginally significant in the ROI analysis.

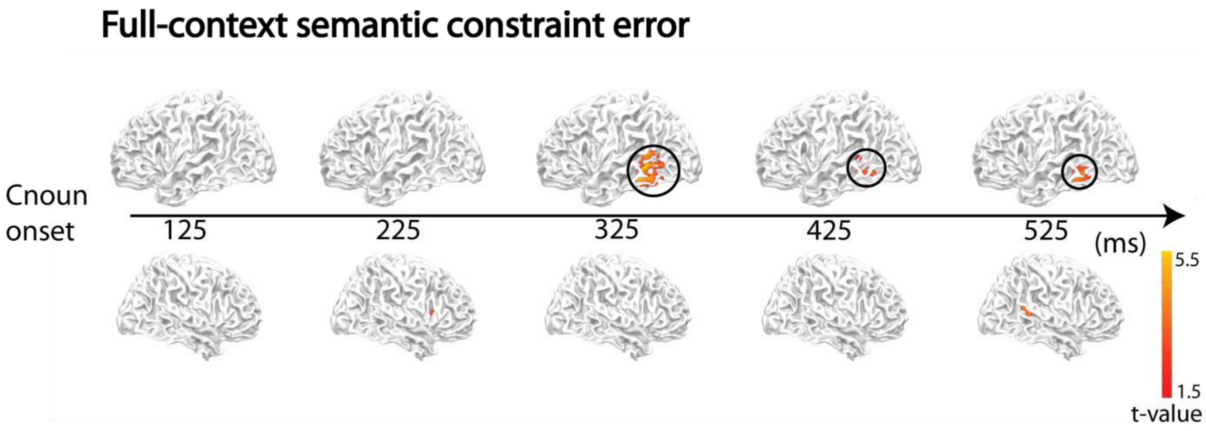


Figure 3-4b: the searchlight clusters of a full-context semantic constraint error. Any clusters inside the bold circles are consistently observed in the ROI analysis.

ROI analysis

Despite the advantage of the searchlight analysis to investigate the spatiotemporal dynamics of neural computations involved in incremental speech comprehension, the searchlights covering the entire 3-D brain over time did not show any significant effects after the multiple comparisons correction. As a next step, it was hypothesized that the effects will be well localized into a set of anatomically defined regions. In this way, the spatial patterns in the neural activity were defined by each of the anatomical ROIs, consisting of 15 different regions in each hemisphere parcellated from the language network defined by Harvard-Oxford cortical atlas. Then, the multiple comparisons are corrected for the number of statistical tests over time using the same approach as described in section 3.2 simply by replacing searchlights over space with every anatomical ROI. The ROI analysis followed the exactly same parameters and procedure as the searchlight analysis described in section 3.2 above.

Following on from the searchlight analysis, this section reports statistically significant results from the exploratory ROI analysis organised as follows. Sections 3.5.1(a) and 3.5.1(b) present the constraint modelling of the computed representation of the SNP + Verb context. Sections 3.5.2(a) and 3.5.2(b) present the surprisal-based probes of the neural consequences of these constraints for the complement phrase. In addition, Section 3.5.3 shows the activation of the semantic contents of the subject noun, a primary source of the full-context semantic constraints. See Table 2 below, showing the main questions addressed in this results section.

Table 2: main questions tested at different sub-sections in the results section. The summary figure 3-10 also helps to address these questions

Question being tested	Section
Q1) What are the linguistic bases of predictive computations?	3.5.1. and 3.5.2. show the significant model-fits of constraint and error with both lexical (single-word) and contextual (full-context) bases
Q2) Are syntactic constraints activated prior to the activation of semantic properties in order to	3.5.1. shows the earliness of syntactic (SCF) and

enable early phrase structure building before constraining the lexical-semantics?	semantic constraint effects
Q3) Do listeners utilize these constraints to guide the interpretation?	3.5.2. shows significant effects of syntactic and semantic error

3.5.1(a) Incremental representational constraints: Semantic

(i) Full context models:

To determine the timing and location of the activation of probabilistic semantic/pragmatic constraints on the complement nouns, I tested for model fit of the full context semantic constraint RDM in an epoch aligned to verb onset and extending 500ms before and after this alignment point (see Epoch V1 in Figure 3-3). Restricting the analyses to ROIs in an extended language system mask, I saw significant model fit in several left and right hemisphere (RH) ROIs (see Figure 3-5). Model fit was seen first in RH anterior portion of superior temporal gyrus (RaSTG) followed by temporal pole (RTP). These effects were followed by left BA47 after the verb-onset. More specifically, these semantic effects (see Figure 3-5) were first seen around the onset of the subject noun in RaSTG from 390ms before the verb-onset lasting around 60ms (RaSTG: p=.031 at -370ms), most likely reflecting the initiation of semantic constraints on the complement phrase based on early context information. This early effect was followed by a stronger peak in RTP from 100ms before the verb-onset (RTP: p=.018 at 0ms). A later peak was found in L-BA47 from 310ms after the verb-onset lasting for about 100ms (L-BA47: p=.022 at 330ms). These effects are all supported by the searchlight results showing the early right temporal effects before the verb-onset which transition into LH after a verb is recognized (see panel (A) in Figure A1).

I further investigated the effect of full-context semantic constraint more specifically with respect to the target word by changing the alignment point from a verb onset to a complement noun onset. In this complementary analysis, the same RDM was tested in an epoch aligned to the complement noun, extending 500ms before and 300ms after this alignment point (see Epoch CN1 in Figure 3-3). The effect initially emerged around 100ms before the target in the bilateral anterior temporal regions including RaSTG and LaMTG and was followed by later effect in LBA45 from about 120ms after the onset (RaSTG: p=.031 at -90ms; LaMTG: p=.05 at -90ms; LBA45: p=.042 at 150ms). The late effect in LBA45 possibly reflects utilizing the constraint to facilitate the processing of the complement noun.

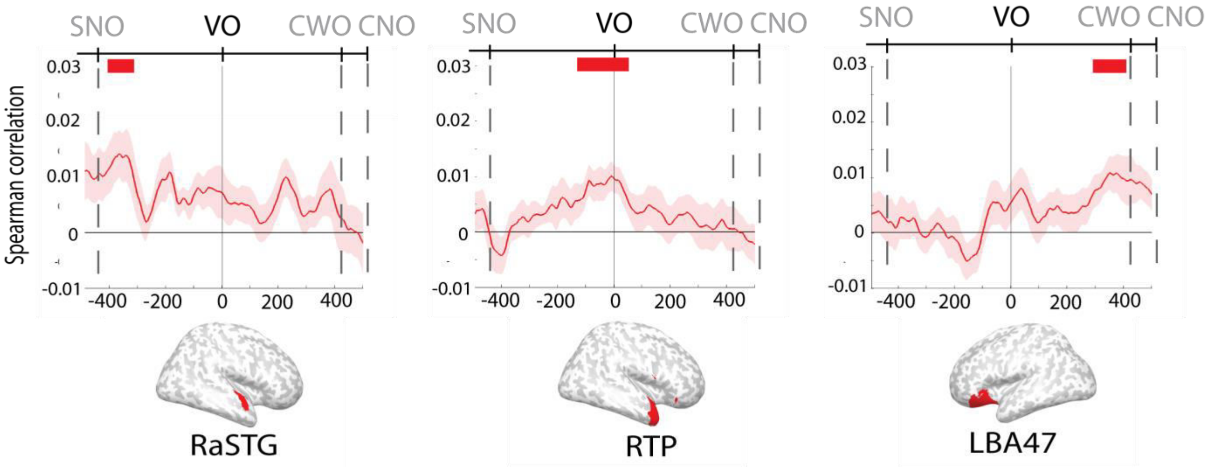


Figure 3-5: ROIs with significant RSA model fits in Epoch V1 for full-context semantic constraints. The Spearman correlation time-courses for full context semantic model fit for the 3 significant ROIs (RaSTG; RTP and LBA47). The time periods of significant model fit are indicated by red bars across the top of each ROI plot. These values are corrected for multiple comparisons across time using threshold free clustering enhancement (TFCE). The peak fit in each ROI (determined by TFCE) is as follows: RaSTG at -370 msec, RTP at 0 msec, and L-BA47 at 330msec (see main text for p-value). The VO alignment point is marked by the black horizontal line in each figure whereas the dotted line reflects the estimated onset of a particular word from the average word durations across trials (indicated by the abbreviation above the plot in grey). The shaded ribbon on either side of the red correlation line indicates standard error across subjects. SNO = subject noun onset, VO = verb onset, CWO = complement function word onset and CNO = complement content word onset.

(ii) Verb alone models:

I tested the semantic probabilistic constraint models based on the verb-alone data in Epoch V2 and CN1 (see AP-V and AP-CN in Figure 3-3) along the same lines as the tests of the full context models reported in Figure 3-3. There were no significant model fits in any ROI for both the WordNet and the topic based models in both of these epochs.

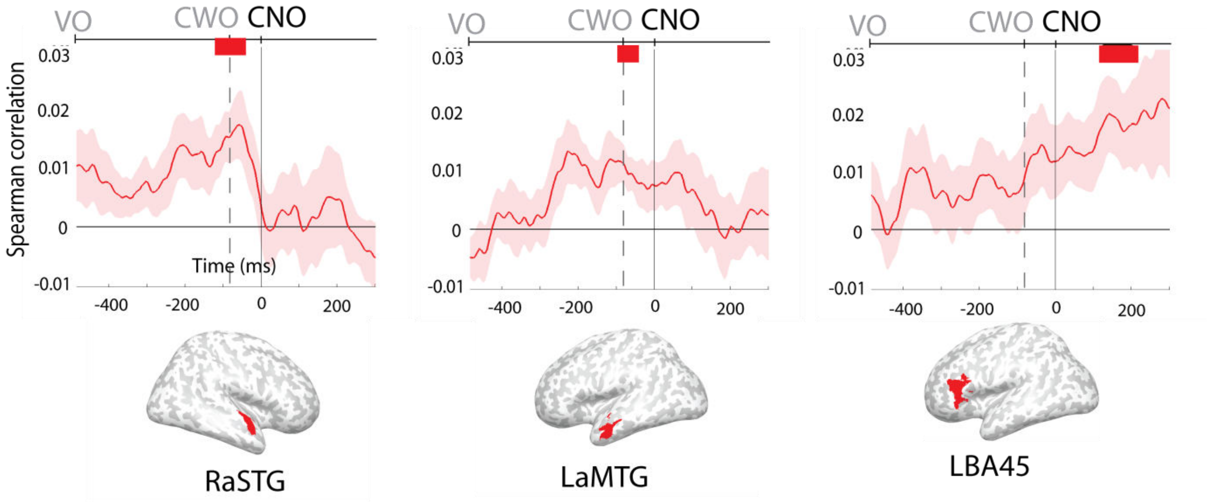


Figure 3-5(a): ROIs with significant RSA model fits in Epoch CN1 for full-context semantic constraints. The Spearman correlation time-courses for full context semantic model fit for the 3 significant ROIs (RaSTG; LaMTG and LBA45).Each peak fit in ROIs is as follows: RaSTG at -90ms, LaMTG at -90ms and LBA45 at 150ms. Other details and annotations are as described in Figure 3-5.

3.5.1(b) Incremental representational constraints: Syntactic

(i) Full-context models

To determine the time-course and neural location of the effects of the probabilistic syntactic constraints generated by the SNP+verb on the complement phrase, I tested for syntactic constraint model fit in Epoch V1 (see Figure 3-3), beginning 500ms before verb-onset and ending 500ms after verb-onset, thereby including both the subject noun and the verb. Significant model fit was found only in L BA44, from 160 msec to 300 msec after verb onset (peaking at 260ms after VO (p=.015; Figure 3-6A)). These results suggest that when a verb is preceded by its subject noun, the constraints generated by the verb's subcategory information become available early in the processing of the verb, only after 170ms has been heard (the average verb duration was 420ms). Note that this effect seems to arise even before VO which may suggest that a subject noun also (weakly) contributes to constraining the complement structure but this will not be further discussed as none of the time-points before VO were significant.

(ii) Verb alone models

To determine the contribution of the verb to the generation of syntactic constraints, I tested for model fit to syntactic constraint models generated by the verb alone. Using Epoch V2 (see Figure 3-3), covering a 600 ms time-period from VO, we found significant SCF model fit initially in LMTG (p=.011 at 230ms after verb onset), appearing slightly later in LHG (p=.02 at 330ms) and peaking in L-BA44 around 350ms-500ms after verb-onset (p=.008 at 500ms) which extended briefly into the function word (see Figure 3-6B). These results are highly consistent with the searchlight results (see panel (B) in Figure A1) and show that the complement structure constraints generated by a preceding verb are activated rapidly as soon as the verb is recognized which last approximately until the function word is identified. Since the function word strongly determines the actual complement structure in the sentence, these results indicate that the verb's syntactic preferences are activated early during the verb's processing and last until these constraints are confirmed or rejected by the actual syntactic structure of the continuing sentence.

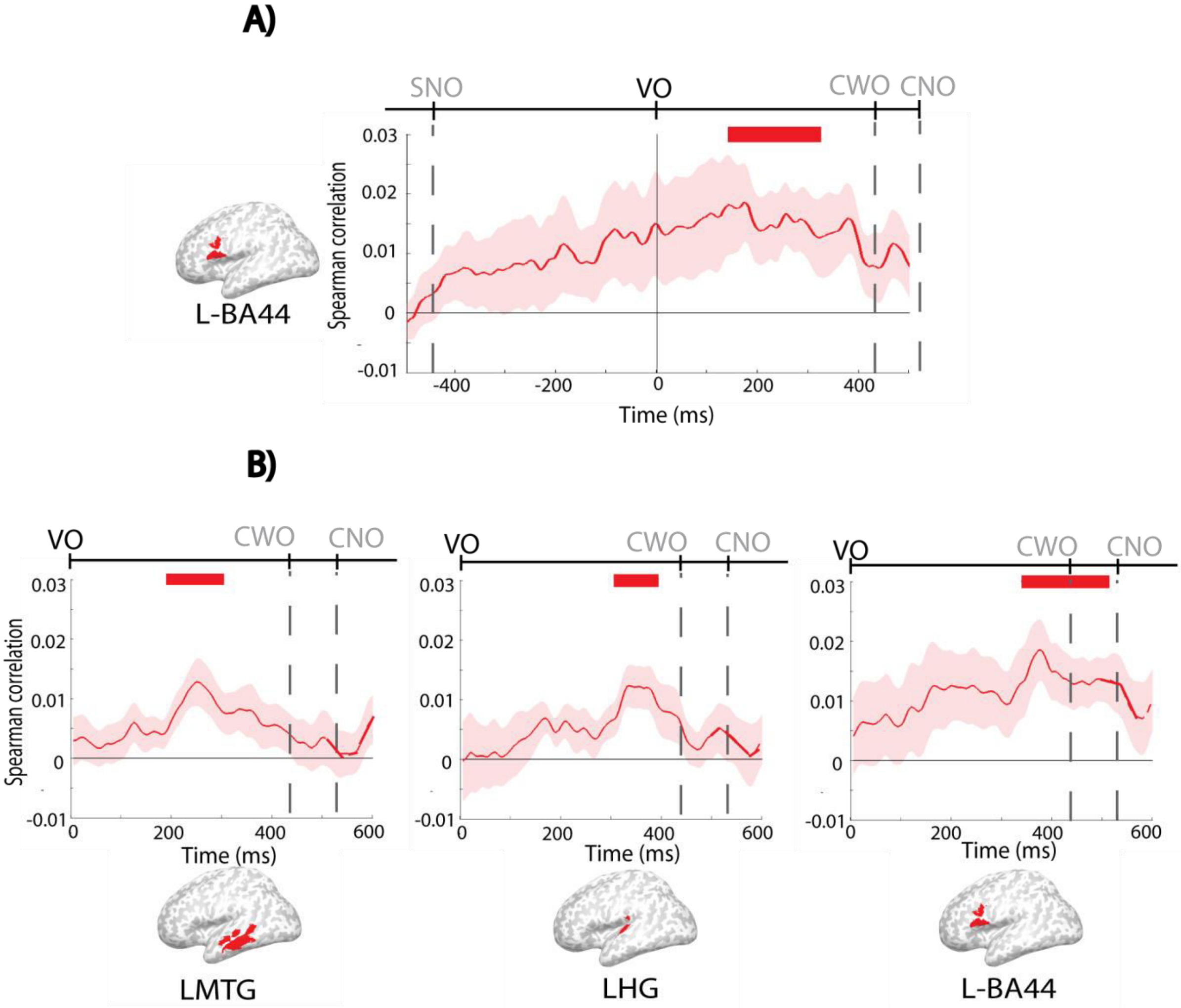


Figure 3-6: Significant RSA model fits for full-context (A) and verb-alone (B) syntactic constraints. Panel A): The Spearman correlation time-course for the full-context syntactic constraint model fit in L-BA44. Red bars indicate TFCE-based significant model fit. Peak fit is at 310 msec. after verb onset. Panel B): Time-courses of model fit for the verb-alone syntactic constraint model in 3 significant ROIs. Peak fit in each ROI is as follows: LMTG at 230 msec, LHG at 330 msec and L-BA44 at 500 msec. Details and legend as in Figure 3-5

3.5.2(a) Integration effects: Semantic

(i) Full-context model

The second set of analyses focused on the processing relationship between the incremental constraints and the interpretation of the complement phrase. I tested for effects of the difficulty of integrating semantic constraints based on the prior SNP + verb context with the semantics of the upcoming complement noun using the full context semantic constraint error RDM in Epoch CN2 (aligned to complement noun onset (CNO) – see Figure 3-3). This RDM captured the semantic distance between the semantic constraints generated by the full context and the semantics of the actual direct object nouns in each sentence. As noted above, I only included the 100 trials with direct object sentences for this analysis.

As Figure 3-7 shows, a significant full-context error effect emerges around 280 msec after complement noun onset in the LpMTG ROI (p=.002 at 340ms and p=.002 at 455ms). This is around the time the complement noun is likely to be recognised – at 323 ms (SD 77) based on uniqueness point estimated from the CELEX database (Baayen, Piepenbrock & Guilikers, 1996). Especially, the early peak was well replicated by the searchlight analysis showing an extensive cluster in the left posterior temporal regions around 325ms after the onset, lasting until around the end of this epoch. Consistently, the effect in LpMTG remained strong until around 500 msec after complement noun onset with a second peak (p=.005) at 450 msec and declined afterwards (the effect becomes non-significant soon after 600 msec post noun onset). Figure 3-7 also includes, for comparison, the non-significant semantic error effect for the topic-based verb-alone semantic surprisal model.

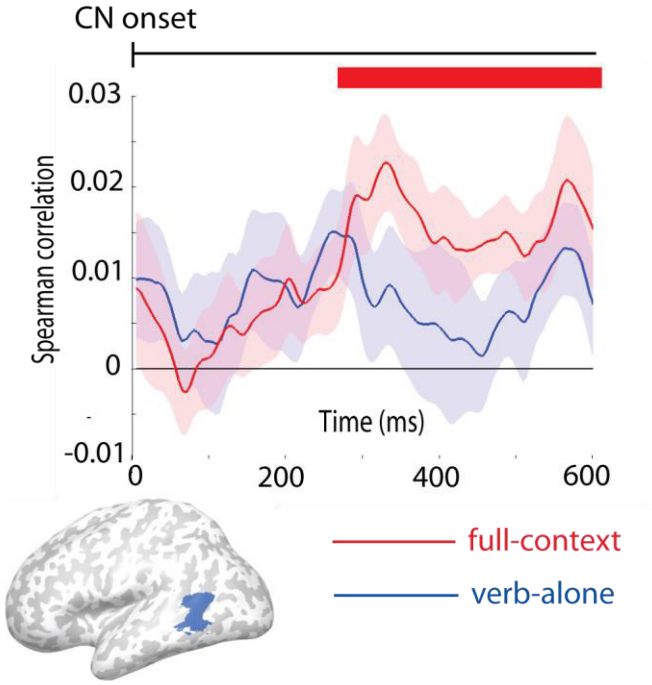


Figure 3-7: RSA model fit for full-context and verb-alone semantic constraint error. The Spearman correlation time-course for the semantic surprisal model; The red line shows the significant and sustained model fit in L posterior MTG (LpMTG) for the full-context surprisal model, peaking at 340 ms, (close to the 323 ms uniqueness point). For comparison, the blue line shows the non-significant verb-alone semantic surprisal model fit, peaking at 265 msec.

3.5.2(b) Integration effects: Syntactic

(i) Full context model

I next looked at the fit between the syntactic constraints imposed by the SNP+verb and the upcoming speech by testing for syntactic surprisal (Figure 3-8), focusing on a testing epoch aligned to the onset of the complement phrase function word (see Figure 3-3, Epoch CFW). The function word constrains the possible syntactic frames that can follow the SNP + Verb context, and is the earliest point at which prior syntactic constraints could have an effect. For the full-context syntactic surprisal model I found only a weak, marginally significant model fit in LBA45 (p=.071) at 175ms after function word onset, as listeners were hearing the complement noun (Figure 3-8).

(ii) Verb-alone model

I tested for model fit of the verb-based SCF surprisal RDM in the same epoch aligned to function word onset (Epoch CFW in Figure 3-3). We found significant model fit only in L-

BA45 (p=.01 at 190ms; see Figure 3-8) peaking around 200ms after function word onset. This effect occurred soon after the syntactic constraint effect ended in L-BA44 (Figure 3-6).

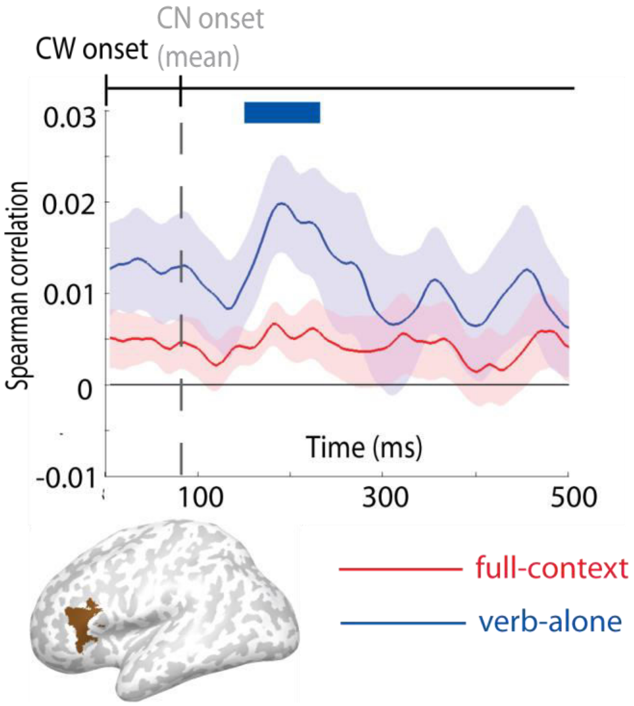


Figure 3-8: RSA model fits for full-context and verb-alone syntactic surprisal. The Spearman correlation time-courses for the verb-alone syntactic surprisal model fit (in blue) for the significant ROI: L-BA45. For comparison, the full-context syntactic surprisal model fit, which was marginally significant (p=.071), is also shown in red. Peak fits for the full-context and verb-alone effects are at 175 msec and 190 msec post CN onset respectively.

3.5.3. Activation of the semantic contents

In this additional analysis, I tested the semantics of the subject noun, as defined by co-occurrence properties at the same epoch in which the full-context semantic constraint model was tested (Epoch V1 in Figure 3-3). The subject noun is a primary source of the full-context sentential constraints that first constructs the event structure by informing the likely thematic role that it is involved in. As expected, the model showed significant correlation with the pattern of neural activity within a window of the SNP in the RH posterior temporal regions unlike the effects of the full-context semantic constraint which involved the anterior temporal and inferior frontal regions in a more distributed time window across the subject noun and the verb. More specifically, the effects emerged in a number of right posterior temporal areas around 380 msec before the verb-onset generally having two peaks in time (RHG: p=.008 at -

340ms and p=.011 at -180ms, RpSTG: p=.003 from -330ms to -300ms and p=.003 at -110ms, RMTG: p=.013 at -360ms, RpMTG: p=.033 at -130ms, RpSMG: p=.022 at -330ms, RaSTG: p=.039 at -290ms, RaITG: p=.035 at -120ms, LBA44: p=.022 at -100ms). Figure 3-9 shows a set of correlation time-courses associated with each of these regions. In contrast to the full-context constraint effects (Figure 3-5), all significant effects of a subject noun semantics are specifically within the window of the subject noun before the onset of the main verb.

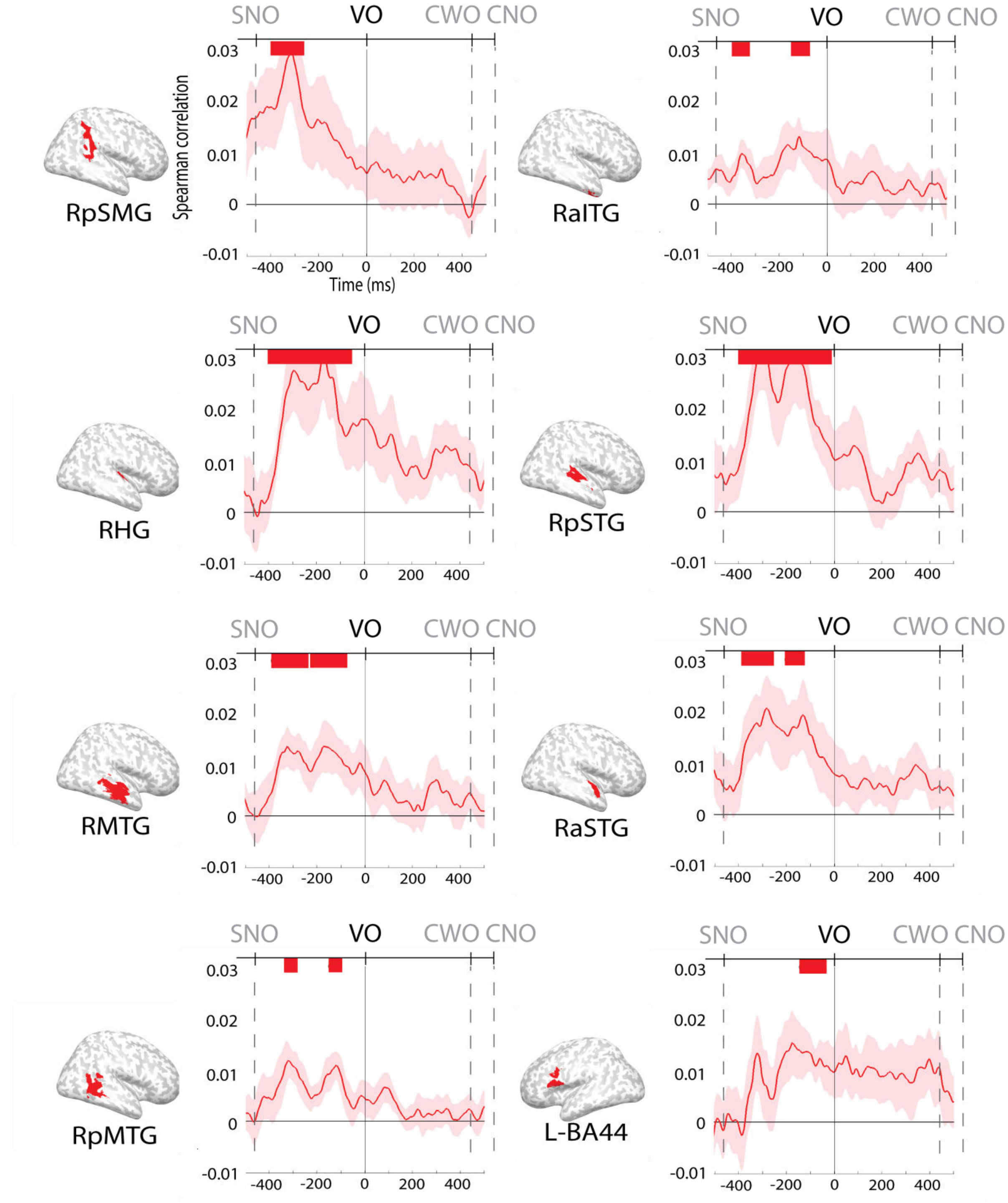


Figure 3-9: Significant RSA model fits for the subject noun semantics. The Spearman correlation time-courses for the subject noun semantics model fit for the 8 significant ROIs. The peak fit in each ROI is as follows (RMTG at -360ms, RHG at -340ms, RpSTG from -330ms to 300ms, RpSMG at -330ms, RaSTG at -290ms, RpMTG at -130ms, RaITG at -120ms and L-BA44 at -100ms). Details and legend are as in Figure 3-5.

3.6. Discussion

The goal of this study was to determine the types of neural computations involved in activating syntactic and semantic contextual constraints in real time as listeners hear spoken sentences, and how these constraints function to facilitate the rapid integration and interpretation of the syntactic and semantic properties of the upcoming speech input. Listeners heard sentences consisting of a subject NP, a verb, and a complement phrase, where the subject NP and the verb varied in the cumulative probabilistic constraints they generated on the upcoming complement phrase and its constituents. I tested for the timing and neural location of these computations by recording real-time brain activity using EMEG and analyzing these spatiotemporal neural activity patterns across an extensive set of bilateral fronto-parietal-temporal regions, using probabilistic models of different types of incremental constraint and their effects on the integration of upcoming words. The two different sets of analyses based on ROIs and searchlights were highly consistent, showing that the clusters (despite not being significant for searchlight) in different ROIs can be reproduced using more spatially fine-grained searchlights, except some relatively weak and focal effects such as the syntactic constraint error in LBA45.

3.6.1. Early RH computation of incremental 'event-structure' representations

Interpretive semantic constraints on the upcoming complement phrase start to be generated early in the sentence, as the subject noun is heard and are maintained until the complement content word is heard (Figure 3-10). These incremental 'full-context' semantic processes were seen in a bilateral fronto-temporal network, in contrast to the effects of syntactic constraints which involved only a LH fronto-temporal network. Such predictive processes were first observed in right anterior temporal regions in contrast to the effects of a lexical-semantic activation of a subject noun which mainly involved right posterior temporal regions. These regions are typically associated with the access of the semantic properties of words from speech (Hickok & Poeppel, 2004; Kocagoncu et al., 2017) and reflect the distributional semantic 'topics' that we captured in our computational models. Further, RaSTG was involved in both representing the early constraint and activating the lexico-semantic information of a subject noun that is currently being heard which, in most cases, is the primary source of the semantic constraints on the complement (the late transient LBA44 activation may hint some constraining processes on the upcoming verb phrase which requires the lexical-semantic information of the subject noun). Therefore, this region may bridge between the general lexical semantic information (centred on RHG/RpSTG) and more specific predictive information (centred on RTP). This effect in RTP was more sustained and extended into the verb. Temporal regions in the RH are thought to be involved in the activation of coarse-grained semantics with larger and more diffuse semantic fields (Beeman et al., 1994; Jung-Beeman, 2005). The earliness of these effects suggest that as soon as listeners hear a subject noun phrase they start to construct semantic-pragmatic models of 'event-level' scenarios of what is likely to be being talked about (Johnson-Laird & Byrne, 2002; Elman, 2011), providing cumulative constraints on the semantic-pragmatic properties of the upcoming speech.

A late effect of the semantic constraint model was observed in L-BA47 several hundred milliseconds into the verb after these RH effects disappeared. Similarly, a late effect was also observed in L-BA45 appearing at around 100ms after the target (complement noun) onset. These frontal model fits, following on from activity in temporal cortex, may reflect the role of frontal cortex in processes of semantic selection and control as more of the verb is heard (Kan & Thompson-Schill, 2004; Thompson, Henshall & Jefferies., 2016; Zhuang et al., 2012). Further, this anterior portion LIFG (LBA44/47) is commonly observed for semantic processing during sentence comprehension (Rodd et al., 2005; Ye & Zhou, 2009; Price, 2010) and is a locus of ambiguity resolution and unification (Hagoort, 2013).

The bilateral networks involved in constructing these incremental interpretative representations are closely linked to LH processes, since the effects of semantic constraints on integrating the complement noun (CN) into the cumulative 'full-context' representation are left-lateralised to LpMTG about 250 msec after the onset of the complement noun (see Figure 3-7). This apparently LH-specific process was detected by a semantic surprisal RSA analysis which captured the degree to which the semantic properties of the noun match or mismatch with the prior semantic context. This effect is broadly similar in its timing to the N400 response (Simos et al., 1997), also thought to reflect the difficulty of accessing the lexical information of the target and of integrating it into the prior context (Lau et al., 2008).

Unlike the model fit for the full-context semantic models in bilateral fronto-temporal network, and the significant semantic surprisal model fit as the complement noun is being recognised, the verb-alone models of semantic constraints - based solely on the lexical properties of the verb preceding the complement phrase - produced no effects in any ROI. This result, taken together with the results of the full-context semantic model, suggests that in normal language comprehension the semantics of the verb interact with the likely scenarios generated by the semantics of the agent (SNP). For example, without first hearing the phrase "*the giant* *crocodile*" in the sentence *"the giant crocodile killed the ..."* it would be difficult to constrain what could be killed: "crocodile" is likely to kill "deer/boar/fish" whereas "householder" is likely to kill "fly/roach/spider". In line with this argument, I observed significant model-fit of the lexical-semantics of a subject noun which presumably interacts with the event representation to generate constraints. This suggests that the incremental integration of sentential semantic constraints cannot be accurately modelled by individual lexical representations alone and most plausibly reflect the initial activation of a broad range of semantic information activated when the SNP is heard. This information combines with the upcoming verb to constrain the semantics of the complement noun. These findings are consistent with the view that the semantic information activated when a word is heard involves both lexical semantics and world knowledge which are activated simultaneously (Tyler & Marslen-Wilson, 1977; Marslen-Wilson et al., 1993; Kamide et al., 2003; Nieuwland & Van Berkum, 2006) and involve the same brain regions (Hagoort et al., 2004).

3.6.2. LH computation of bottom-up lexically-driven syntactic constraints

In contrast to the early and extensive RH model fit for full context semantic constraints, model fit for the full-context syntactic constraint model only became significant later in the sentence - from 160ms to 300ms after the verb-onset - and only in the LH, in LBA44 (Figure 3-10). The timing of this late, transient and limited effect of the subcategorisation constraints on the complement phrase tells us that the event representation (or message-level semantics) starts to be constructed even before the associated syntactic structuring emerges (although the full-context syntactic model-fit started to gradually rise after the subject noun onset, the full-context semantic effects were already significant during the subject noun). These results are consistent with the view that the contextual semantics/pragmatics guide upcoming syntactic (Tyler & Marslen-Wilson, 1977) and semantic interpretations (Altmann & Steedman, 1988; Altmann & Mirkovic, 2009; Marslen-Wilson et al., 1993). It is notable that we see the significant effects of both the semantic and syntactic constraint models at around the same time, but in different brain regions – LBA47 for semantic constraints and LBA44 for syntactic constraints - providing evidence for the parallel computation of both syntactic and semantic constraints as relevant information becomes available.

The timing and location of the full context syntactic effect differed from the effects of the syntactic constraint model based on the verb's lexical properties alone (see Figure 3-6). Here the effects emerged slightly later and were located in a more extensive set of regions (all restricted to the LH) including LMTG peaking at 250 msec post verb onset, followed by transient effects in LHG and robust effects in LBA44 from 400msec which persisted into the complement word, replicating previous results for a verb-alone analysis on an independent data-set (Tyler et al., 2013).

This pattern of results suggests that effects of the full-context syntactic constraints seen in LBA44 reflect the incremental integration of the prior context with the lexical constraints generated by the verb activated in LMTG. These lexical constraints are evaluated against the actual syntactic structure in L-BA45 very quickly from 170ms to 230ms after the onset of the complement noun. A similar pattern was observed for the full-context surprisal effect in L-BA45 in this window but only to a weaker extent. In contrast to these relatively local effects for syntax, the semantic constraints reflected in the full-context semantic model including the subject NP predominantly guides the semantic interpretation of the complement phrase.

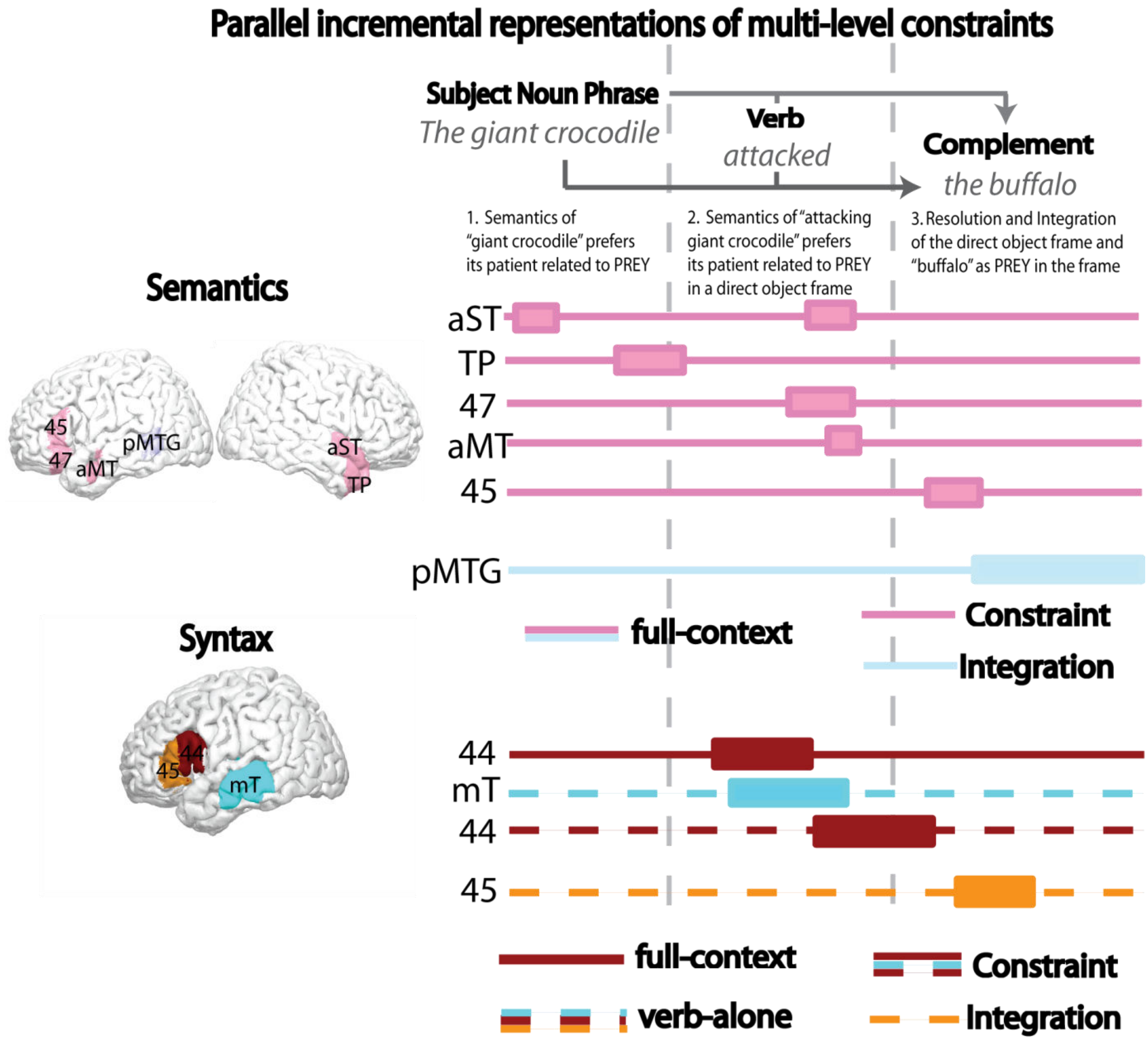

Parallel incremental representations of multi-level constraints
Subject Noun Phrase
The giant crocodile
Verb
attacked
Complement
the buffalo
1. Semantics of "giant crocodile" prefers its patient related to PREY
2. Semantics of "attacking giant crocodile" prefers its patient related to PREY in a direct object frame
3. Resolution and Integration of the direct object frame and "buffalo" as PREY in the frame
Semantics
45
47
aMT
pMTG
aST
TP
aST
TP
47
aMT
45
pMTG
full-context
Constraint
Integration
Syntax
45
44
mT
44
mT
44
45
full-context
verb-alone
Constraint
Integration

Figure 3-10: Summary of results in the bilateral language network. RSA effects of syntactic and semantic constraints and integration during language comprehension. The effects of full-context semantic constraints and integration are summarized in pink and pale blue respectively. The bottom panel shows the effects of full-context (in bold) and verb-alone (dotted) syntactic constraints (in dark red and cyan) and integration (in orange). The relative timing of each effect is shown by a bar(s) on the line that represents each region.

3.6.3. WordNet-MDL based approach of modelling semantic constraints

Propagating the lexical constraints to the WordNet conceptual space directly produces the constraints in the hierarchically organized semantic senses. Although it provides the representation of constraints in the well-established semantic space consisting of clearly interpretable senses, modelling the entire semantic space consisting of 117,000 senses is just as expensive as modelling the lexical space. Further, the vast majority of senses are redundant in terms of representing the constraints because 1) many senses will have a frequency value of zero, especially those in the bottom leaves and 2) these senses are all hierarchically organized which means that any sense in the upper hierarchy (i.e. hypernym) becomes redundant if all of the lower-level senses that belong to it (i.e. hyponyms) are properly represented. As described above in the section 2.5.2(a), these are the reasons for finding an optimal cut which can maximally preserve the original variability in the entire semantic space with a minimum number of semantic senses.

However, I did not find any significant effects of the WordNet-MDL model consistent with the topic model based on the document-topic (DT) distribution. These results strongly converge to the claim that the semantic constraints on the complement depends heavily on the entire context including a preceding subject noun, consistent with the prediction accounts of language comprehension. Although this provides an explanation to why this model failed to capture the temporal patterns of activity in any of the ROIs, there are some other potential limitations that need to be taken into account.

First, it does not fit very well in the RSA analysis framework. The WordNet-MDL model is based on the mean optimal cut across 50 different verbs where each verb has a specific optimal cut which could be vastly different from the other verbs. For example, the highly constraining verb “suffer” has very specific preference of senses including “collapse”, “disease”, “frostbite”, “bleeding”, “crash”, “pathology”, “infection”, “calamity”, “disorder”, “misfortune”, “distortion” and so forth. These senses are optimally represented (through the tree-cut MDL optimization described in the section 2.5.2(b) in Chapter 2) for a particular verb “suffer” and, as a result, the mean optimal cut summarizes these senses into a hypernym sense, losing the specificity of representation. This is inevitably the case for the other constraining verbs in the stimuli. Despite 142 senses being optimally represented in “suffer”, the recursive evaluation scheme to find the mean optimal cut across 50 verbs in our stimuli only returns the 15 very general senses in the end. Other than the less constraining verbs such as “want” and “try” which only have 7 optimally represented senses (“causal agent”, “object”, “matter”, “process”, “abstraction”, “thing1” and “thing2”), this algorithm likely have lost its specificity to the constraining verbs because it forced all 50 different verbs to be represented by a common set of semantic senses in order for them to be comparable.

Further, the distance between representational geometries of two vectors of constraints is highly sensitive to the level of representation in the hierarchical semantic space. For example, consider two different verbs “eat” and “drink”. In a simplified semantic tree in which the “food” synset consists of “foodstuff” (preferred by “eat”) and “beverage” (preferred by “drink”), the optimal cut may include either “food” or “foodstuff” and “beverage”. If the optimal cut is at the hypernym “food”, the distance will be very close as both verbs strongly prefer “food” to be its complement. In contrast, if the optimal cut is at the hyponyms “foodstuff” and “beverage”, these two verbs will be highly dissimilar. Therefore, finding the mean optimal cut across 50 different verbs could lead to the loss of distinction between the semantic constraints in the RSA framework. Additionally, although the optimal cut is found in a way that maximally preserves the original variability in the entire semantic space with a minimum number of parameters (synsets), it cannot account for the large individual variability in the representation of the conceptual semantic space (e.g. a butcher’s constraint could be different from a carpenter’s constraint after hearing a verb “cut” in a sentence). In summary, with RSA which is sensitive to the dimensions of representation, modelling the semantic constraint in the hierarchical semantic space can be difficult.

3.6.4. Conclusion

The results of this novel study show that the brain constructs multi-level probabilistic constraints as soon as the relevant information becomes available and these constraints are adapted and carried forward throughout the sentence via rapid incremental integration.

Semantic information is accessed first serving to create broad event structures which constrain the upcoming speech and is underpinned by temporal and frontal regions in the RH. Both syntactic constraints and the computations involved in integrating the complement noun into the prior syntactic and semantic context are strongly left-lateralized. These results are consistent with models of language processing which emphasise the important contribution of semantic context over syntactic principles (e.g. minimal attachment) in generating on-line multi-level constraints (Taraban & McClelland, 1988).

Chapter 4: Decoding the internal representation of a predictive machine and testing its relation with the neural computations

Incrementality is a fundamental aspect of speech processing, involving a wide range of complex computations that interpret sequentially unfolding inputs based on the preceding context and integrate them into structured and meaningful phrases, sentences and discourses. Although the previous study in Chapter 3 showed how a preceding context could constrain the upcoming input at syntactic and semantic levels, it did not directly address the incremental changes in the state of the brain which alter the constraint on the sequence of words. Not many studies have looked at the spatiotemporal properties of the complex neurobiological systems that support these dynamic, word-by-word transformations. In this study, I use a state-of-art computational model based on the connectionist theory of cognition which allows us to investigate the incremental alteration of the internal state and its relation to the output prediction. Using this model, the dynamic patterns of time-sensitive neural activity related to incremental, word-by-word computations are thoroughly investigated. Further, validity of this connectionist model as a descriptive measure of human incremental speech comprehension is tested by comparing the results with the previous study in Chapter 3 using behavioural and corpus-based models.

Following on from the previous chapter showing the predictive nature of human speech processing, this chapter further investigates the computation of constraints and the nature of its representation using a well-trained connectionist model, designed to address various incremental computations during speech comprehension. Within the predictive framework of speech comprehension, the debate continued regarding the level of abstraction of the preceding context for computation of the constraints. Consistent with many connectionist views, the sequential processing account argues that the word-level statistical information is sufficient for explaining the majority of computations involved in incremental speech processing. In contrast, the hierarchical processing account claims that tracking the hierarchical constituent structure (i.e. the abstracted syntactic information) is essential for computation of linguistic constraints and easily explains some complex language comprehension phenomena such as long distance dependencies. In other words, the debate is whether human speech comprehension is driven by decomposing the hierarchical structure of a sentence through abstraction or by understanding the statistical relation between each word in a sentence. A key word that distinguishes these two accounts is “abstraction” such that one emphasizes the abstracted word-category information to constrain a syntactic position of the word in a sentence for phrase structure building unlike the other arguing that the word-level statistics is sufficient to explain a variety of speech comprehension processes.

One of the main goals of this chapter is to test if statistical information of words in context is sufficient to guide the interpretation of an upcoming word in humans without any explicit definition of syntax or any syntactic supervision during training. In particular, I aimed to investigate the representational contents in the state of the connectionist network and further explore the dynamic nature of computations in the network by modelling its incrementally changing representation throughout a sentence. In this way, it shows how well the connectionist account of computation explains the neural computations involved in generating and utilizing the constraints in a predictive framework.

4.1. Sequential vs. hierarchical processing accounts of human speech comprehension

A recent research article in Nature Neuroscience again sheds light on the importance of syntax in understanding speech (Ding, Melloni, Zhang, Tian & Poeppel, 2016). In a cross-linguistic study between Chinese and English, the authors showed clear peaks in neural responses at the frequencies at which the stimuli are processed at different levels. In particular, they observed 3 clear peaks in the frequency spectrum of neural responses at 1Hz (a sentence presentation rate), 2Hz (a phrase presentation rate) and 4Hz (a syllable presentation rate) by presenting a sentence consisting of two phrases each of which contains two syllables with a presentation rate of each syllable for every 250ms (e.g. “*new plans gave hope*” consisting of NP and VP). This pattern of results, however, was not observed when listeners did not understand the language. For example, the cortical activity of English speakers when listening to Chinese stimuli only showed entrainment to the syllabic rhythm at 4Hz. Consistent with other neuroimaging studies of artificial syntax showing that statistical cues are not necessary to trigger neural tracking of the structure in a sequence, they interpreted these results as evidence for cortical tracking of hierarchical structures in a sentence and supported the claim that the brain can form representations at various syntactic levels based solely on rules (Ding, Melloni, Tian & Poeppel, 2017).

Nevertheless, an obvious question one has to ask is how applicable these results are in explaining natural speech comprehension in the real-life environment. Nobody speaks at the same rate all the time in real-life communication. Hence, although it can be acknowledged that humans are capable of tracking structure of a sentence based solely on their syntactic knowledge, it doesn't mean that it is necessary to understand a spoken sentence. In fact, processing a sentence more likely depends on the syntactic complexity of it such that a listener's syntactic knowledge may become useful as a confirmatory process involving grammatical analysis on syntactically complex sentences.

Moreover, the pattern of results in Ding et al. (2016) was replicated in Frank & Yang (2018) even when they used a word-level statistics model based on the Skipgram architecture (see Mikolov, Chen, Corrado & Dean., 2013) that knows nothing about such grammatical rules. For each simulated participant, they concatenated the N dimensional column (Skipgram) vectors (where N is randomly sampled for each participant with mean = 300 and SD = 25) into a matrix such that each row represents a time-course of simulated MEG samples for a particular dimension. Each MEG sample was simulated in a way that the column vector only contains Gaussian noise (mean = 0 and SD = 0.5) until t milliseconds after the word onset and the actual information (signal) becomes available only after t milliseconds (the time-point t most plausibly reflect the word's uniqueness point). The signal was added by Gaussian noise to reflect the noise in MEG data. A power spectrum for each row was then computed using discrete Fourier transform (DFT) quantifying the amplitude of a sinusoid in each frequency contained in the row vector and was averaged with the other power spectra across N dimensions. Replicating the original results in Ding et al. (2016) suggested that the cortical tracking of syllabic, phrasal and sentential rhythms can be explained by the lexical information without applying the grammatical rules. This also reflects the possibility that word-level statistics could sufficiently trigger tracking of local phrases, just like it can trigger the learning of syntactic rules (Seidenberg et al., 2002).

In the light of Occam's razor, cognitive science pursues a parsimonious model as a descriptive measure of cognitive processing in humans. The logic is if both simpler and more complex models explain a particular cognitive phenomenon, the simpler model is favoured as a descriptive measure unless the complex model performs significantly better in explaining the phenomenon. Assuming that abstraction requires additional cognitive operations, a non-abstracted model based on the word-level statistics should be favoured (see Frank & Christiansen, 2018). As a side note, it is acknowledged in Frank & Christiansen (2018) that the lexical information captured by distributional models is already abstracted, reflecting syntactic and semantic category information of an input (just like the topic models described in Chapter 2) without engaging syntactic knowledge. However, if such abstracted representation is obtained through years of experience, the lexical information is likely to be represented in processing dimensions optimized through experience without requiring further explicit computational operations for abstraction. This could enable the brain to track the hierarchical structures in a simple sentence commonly used in a daily conversation based on the lexical information.

Following on from this debate, I use a connectionist model designed to process the lexical information in a distributional format in order to generate an accurate prediction of an upcoming word. This connectionist framework provides a transparent predictive machine whose internal state and its relation to the output response can directly be investigated at any particular point in a sentence. Compared to a human brain consisting of billions of neurons (or information processing units), such predictive machine is much simpler in architecture with fewer processing units and has a much more straightforward representation. Comparing how similar the nature of incremental speech processing is between a human brain and a state-of-art predictive machine is an interesting topic that has not been thoroughly investigated in the literature. By decoding the linguistic properties activated in the internal state of a well-trained machine and relating the pattern of activation to the temporal dynamics of neural activity using RSA, this chapter identifies a number of brain regions showing similar pattern of activity as the machine at a time when the multi-level linguistic constraints are activated (see Chapter 3). Further, by modelling the spatiotemporal characteristics of the activity pattern for each ROI using the output prediction of the machine, I evaluate the prediction in the light of the incremental computations in humans during speech comprehension. If computation involved in generating the constraint in the brain is based purely (or partly) on combination of the distributional properties of an input word with its internal representation of the preceding context without any explicit engagement of syntactic knowledge, it is expected to observe significant correlation between the representational geometries of the internal state of the model and the brain in the similar time and regions as shown in Chapter 3.

4.2. Connectionist models in a parallel distributed processing (PDP) framework

Computation in connectionism is grounded in a parallel distributed processing (PDP; Rumelhart, Hinton & McClelland, 1987). Any connectionist model in this framework consists of a set of processing units often organised into input, hidden and output layers. If a layer of units has a distributed representation, it is meaningful to treat the pattern across all processing units within the layer as a whole instead of unit-by-unit interpretation. In some cases, each unit in a layer (usually, an input layer) represents a single, independent concept such as a particular word, letter or phoneme etc. This renders the activation state in the layer to be defined by a binary one-hot representation (i.e. one for the unit associated with an input concept and zeros for all the others). This framework is intrinsically parallel because all units in a layer carry out their computations at the same time.

In this system, processing of an arbitrary input is determined by a pattern of connectivity among different units. Each unit passes its output signal onto the other units through this connectivity pattern. Assuming that the contribution of each unit to the others to which it is connected is additive, the activation state across the receiver units can be computed by the weighted sum of all separate inputs to each of the receiver unit. It is this weight that directly modulates the input to the receiver units; Given the entire weight matrix W, any connection from a unit i to a unit j (i.e. $w_{i,j}$) being greater than zero represents an excitatory connection, any $w_{i,j} < 0$ represents an inhibitory connection and any $w_{i,j} = 0$ represents no connection. Hence, $|w_{i,j}|$ represents the connectivity strength. In a more complex PDP model, the connectivity pattern can be defined by a set of weights W_k for each type of connection k. It is worth noting that this W has some theoretically important implications such as the degree of top-down vs. bottom-up processing in recurrent networks described below.

Given such importance of W, the most important aspect of model training in the PDP framework is to modify W as a function of experience. One of the most influential theories of learning was introduced by Donald Hebb whose basic idea is that the connection $w_{i,j}$ should be strengthened if a unit u_i receives an input from another unit u_j and both are active. This idea can be expressed as:

$$\Delta w_{i,j} = g\big(a_i(t), t_i(t)\big)h\big(o_j(t), w_{i,j}\big) \dots (33)$$

where t represents a particular point in time, $t_i(t)$ is a teaching input to u_i, $a_i(t)$ is an activation value of u_i, $o_j(t)$ is an output signal of u_j and $\Delta w_{i,j}$ represents the change in $w_{i,j}$.

A commonly used variation of Hebbian learning specifies the arbitrary functions g and h as $g\big(a_i(t), t_i(t)\big) = \eta\big(t_i(t) - a_i(t)\big)$ and $h\big(o_j(t), w_{i,j}\big) = o_j(t)$ which makes (33) expressed as:

$$\Delta w_{i,j} = \eta\big(t_i(t) - a_i(t)\big)o_j(t) \ldots (34)$$

where η is a learning rate parameter. A learning rate determines the width of a step when searching for an optimum in the error gradient (i.e. the gradient of a loss function). Moreover, if the teacher t is not available, the learning rule (33) further simplifies to:

$$\Delta w_{i,j} = \eta a_i(t)o_j(t) \ldots (35)$$

These (34) and (35) are the example objective functions of supervised and unsupervised training algorithms which will be dealt in Chapter 4. These formulations determine how a network learns the relevant cognitive patterns in the data to generate a desired response. After training, these weights conduct the computational process of interpreting the input in a semantically coherent fashion by modulating each dimension of representation.

Unlike many other human cognition processes, not only does understanding speech require identifying individual words in a sentence but it also involves interpreting in the context in which each word occurs. The overall results in Chapter 3 show that constraints on an upcoming complement phrase generated by a prior verb alone does not explain much variability in the pattern of neural responses, unless its preceding subject NP is taken into account. These results support the view that each word in a sentence represented in a form of vector embeddings works as an operator that directly alters the current state of the system (Elman, 2011). However, a simple neural network treats every input independently without taking their inter-relation into account when generating an output response. Unlike a simple feedforward network, the vast majority of language networks allow the previous state of the system to alter the way that an input is represented and such altered representation computes an output response at the current state. In this way, the network becomes capable of implicitly representing time in its state and incrementally processes each word as in human speech comprehension.

While adding recurrence renders the network sensitive to the full-context in a sentence, it does not lose the lexico-semantic information of the input word. Elman (2011) showed that such a recurrent network learns to partition its internal representation based on the lexico-semantic properties such that the nouns sharing a similar theme cluster in the internal space.

Therefore, the dynamical properties of the network, allowing a flexible modulation on integrating the lexico-semantic properties with the preceding context, intrinsically reflect their relative importance in generating the optimal response. Moreover, Elman (2011) also illustrated that such dynamical properties (encoded in the weights between units) reflect the syntagmatic relations among phrasal constituents in a sentence. He reported that the geometric representation of the same verb and its argument (e.g. "uses a saw to cut …") at the internal layer varied depending on the preceding context (e.g. "A butcher" vs. "A person") which converged after different patients are presented (e.g. "A butcher uses a saw to cut meat" vs. "A person uses a saw to cut a tree"). This is expected because the states of the network up to "cut" reflect different predictions on the patient, leading to the converging representation once it is identified. In summary, adding recurrence to the network enables the current input word to be interpreted in a syntagmatically coherent fashion with respect to the preceding context encoded in the previous state of the system while preserving the lexical properties of the input in the current state.

How well does this recurrent network perform as a model of incremental speech processing in humans?

In this thesis, I use a more sophisticated variant of the recurrent network, known as long short term memory (LSTM; Jozefowicz et al., 2016) trained through a large language corpus known as one billion word benchmark with nearly 800K types (a.k.a. tokens referring to any lexical units incorporated in a model in NLP) in the vocabulary (Chelba et al., 2013). In addition, it takes character-level embeddings as an input instead of word-level embeddings which improves the flexibility of the network such that its' processing is no longer limited to a fixed vocabulary (i.e. unlike a word, there are infinitely many possible combinations of characters in different lengths which allows the character-level representations to be flexible). The released LSTM is a version called "BIG LSTM – CNN inputs" that showed the lowest perplexity (i.e. highest accuracy in prediction) in Jozefowicz et al. (2016) out of all different variants they tested in their paper. This model is trained using ADAGRAD adaptive gradient descent algorithm described above in 3.2.4 in a truncated back propagation through time (BPTT) framework where BPTT is performed only up to a given time (see 3.2.5). This version of LSTM is publicly available, consisting of two recurrent internal (hidden) layers, each of which contains 1,024 (bottleneck) processing units and an output layer showing the softmax prediction of 793,471 types in the vocabulary (The first internal layer referred to as HL0 in this thesis projects to the second internal layer referred to as HL1 which in turn projects to the output layer for prediction). For more details about the architecture, various training algorithms and their implications in “mental state” of the network, see Chapter 2.

4.3. Decoding the pattern of activation in the LSTM internal and output layers

4.3.1. Sanity checks and methods

Before using this LSTM network model to characterize the spatiotemporal dynamics of neural activity, I explored the nature of information processing in the two hidden layers using a number of linguistic models capturing different aspects of computations involved in human speech comprehension, as illustrated in Chapter 3. These models are the full-context and verb-alone models of constraints as well as a model of the lexical semantic information of an input word which are tested against the brain data and reported in Chapter 3. In this way, I hoped to gain a better understanding of how the network processes an incrementally unfolding sequence of words in a sentence and construct more specific neurocognitive hypotheses for different layers of the network. But, before going into details, one of the key aspects of the LSTM network, recurrence of a theme, can easily be seen from a simple sanity check below.

In order to illustrate that the network is capable of retrieving and applying a recurrent theme when making predictions, it was used to generate a sentence from a given fragment (a simple continuation study). Each word after the fragment was sampled from its output prediction and the sampled word was combined with the fragment to sample a next word until the end of a sentence in the following way:

- *“**The local politician emphasised** that……..*
- *“**The local politician emphasised** that the…….*
- *“**The local politician emphasised** that the issue .....*

For different fragments, the following sentences were generated (a given fragment is marked in bold and a recurring theme is underlined):

- *“**The local politician emphasised** that the issue was the result of political manipulation of the press and the public interest.”*

- ***“The bank manager acknowledged*** *the mistake and notified the FDIC as soon as possible.”*
- ***“The duty solicitor concluded*** *that the claim was not only invalid but also in breach of Article 14 of the European Convention on Human Rights.”*
- ***“The graduate student applied*** *to a university to find out which university he was interested in and then went to a job fair.”*

From these LSTM generated sentences, we can see that the theme of the subject in a given fragment (highlighted in bold) is recurring throughout the sentence, as indicated by the underlined text. This shows that the network is capable of holding the necessary thematic information in its memory so that it can associate the recurring theme in the later part of the sentence to the subject. Again, this is the main advantage of using LSTM architecture, designed to address the vanishing gradient problem through recurrent layers (see Chapter 2).

In order to delve into more details about various linguistic properties being activated by the internal representation of the model at each point in a sentence, I compared the similarity pattern of the internal state at every incremental sequence of words with that of 7 different models of interest, described below, capturing a variety of linguistic properties of incremental computations at five adjacent points in a sentence starting from the subject noun up to a point including the complement noun. For example, in a sentence “*The young man fled the army when the fighting began*”, the five points included the consecutive sequence of words including “man”, “*fled*”, “*the*”, “*army*” and “*when*”. The models of interest included the full-context and verb-alone subcategorization frame (syntax) constraint models (see 2.5.1), the verb-alone WordNet-MDL model capturing the VALEX lexical constraint in the WordNet conceptual space (see 2.5.2(a)), the full-context and verb-alone LDA topic models capturing the co-occurrence relation between a verb and a following noun specifically in a direct object frame (see 2.5.2(b)) and a subject noun and a verb DM models published by Baroni & Lenci (2010) that capture the general co-occurrence properties of the word (see 2.5.2(b)).

Comparing the similarity patterns involved creating a set of RDMs of the LSTM internal activation (see section 3.2) at the five points mentioned above. In the section 3.2.1, I described a number of distance metrics and the properties of each of them. Here, I used the Euclidean distance as a default distance metric to compare the representational geometry of the activation vectors across 1,024 hidden processing units (or neurons) between different trials. Again, this metric is highly sensitive to exact amplitude of each processing unit which is the key information to generate an output prediction via the weighted combination across the processing units in the softmax layer (see 3.2). In contrast, cosine distance was used to model the similarity pattern of the softmax layer consisting of nearly 800,000 units each of which reflects the prediction strength for a particular word in the LSTM vocabulary. Again, the reason for using cosine distance here was to neglect the absolute probability difference for each of the ~800,000 types (i.e. many of the types were not in the human vocabulary) while taking the overall covariance into account. These LSTM RDMs were compared with each of the model RDMs using Spearman’s correlation as described above in 3.2.1. The results are shown in the figures below (Figure 4-1, 4-2).

4.3.2. Results

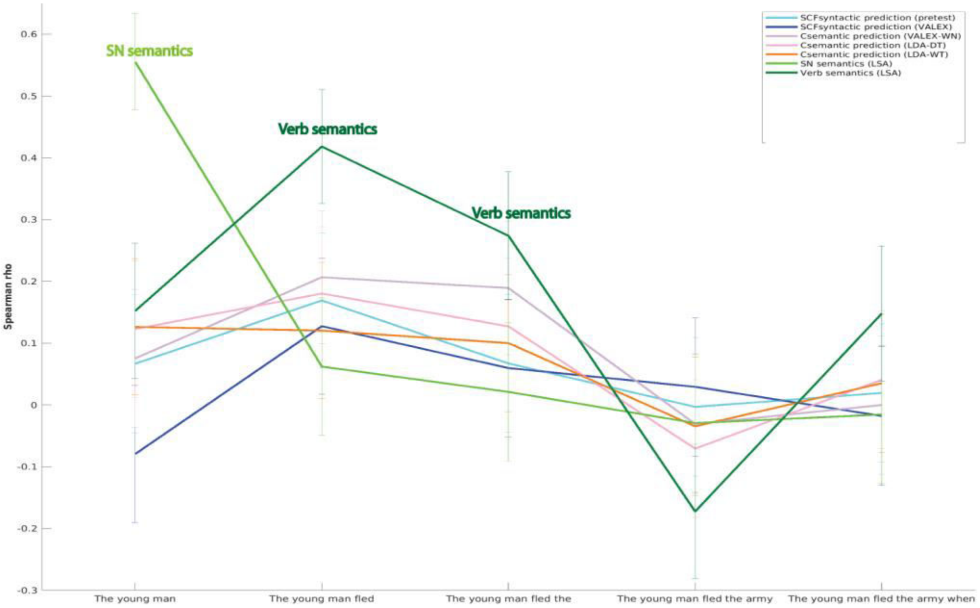


Figure 4-1: A correlation plot of the first Hidden Layer (HL0) with 7 different models of interest at the five adjacent points in a sentence described in the main text. Each line in the plot reflects the correlation time-course associated with a particular model indicated in the legend. The error bars show 95% confidence interval calculated as $\tanh(\tanh^{-1}(\rho) \pm \frac{1.96}{\sqrt{N-3}})$ *where* ρ *is a ranked correlation coefficient and N is the total number of elements in the vectorized RDMs. The inverse hyperbolic tangent (Fisher) transformation on* ρ *renders the sampling distribution to be approximately normal with the standard error of* $\frac{1}{\sqrt{N-3}}$ *and the interval is transformed back to the original scale by applying the hyperbolic tangent function.*

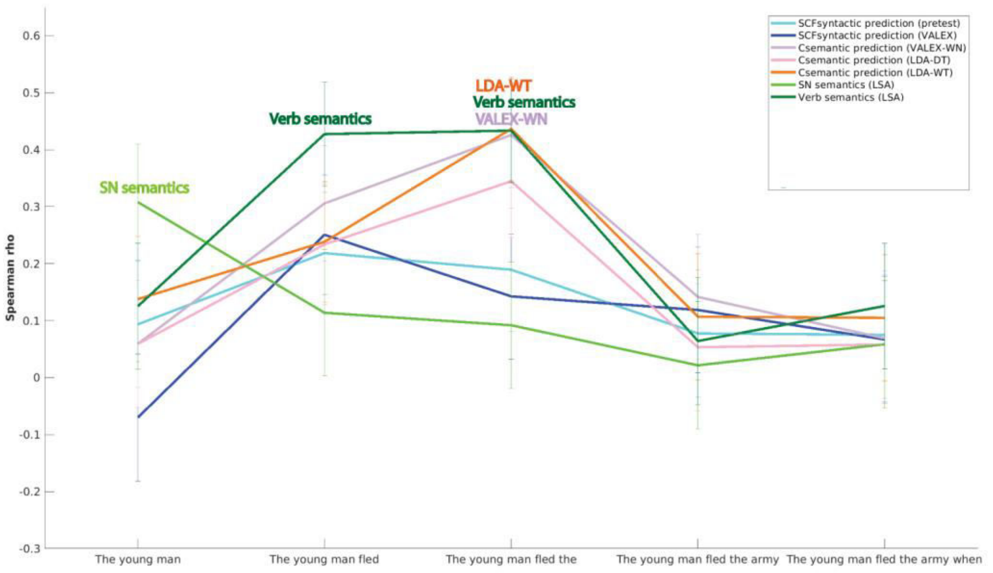


Figure 4-2: A correlation plot of the second hidden layer (HL1). Other annotation details are same as in Figure 4-1.

From Figure 4-1, we can see that the models reflecting the semantic properties of an input word at each point is showing the greatest fit. For example, at the point when a subject noun is revealed "*The young man*", the semantics of "*man*" is activated strongly showing the greatest fit, which immediately declined to the least good fit as soon as the following verb "*fled*" is revealed (light green). A similar pattern was observed for the semantics of "*fled*" which declined immediately after the function word "*the*" is revealed (dark green). From these results, we can infer that the role of the first internal layer HL0 is to activate the semantic information of the input word which will project this information to HL1 for further predictive processing described below.

Next, Figure 4-2 shows a largely different pattern of results. Although the peak effects for the semantics of a subject noun and a verb occurred as they were being heard, the peak effect of the verb semantics did not decline even when the function word in the verb's complement was heard (dark green). Further, the strength of correlations between constraint models and the HL1 state was generally increased where syntactic constraint was consistently activated at the point of a verb (light and dark blue) whereas semantic constraint was activated later, at the point of the complement function word (orange, light pink and purple). As expected, these constraint effects on the complement phrase declined once the actual complement is revealed. From these patterns of results, the information processing in HL1 involves computing and activating constraints on the various linguistic properties of the upcoming continuation including both syntax and semantics.

In order to investigate the information encoded in the output layer of LSTM, the exactly same approach was taken of constructing an RDM from the output vector and of comparing it with 7 different models of interest (see Figure 4-3) as used in Chapter 3. Interestingly, the results showed a different pattern from those related to the internal states. First, the two syntactic constraint models showed strong correlations when a verb is heard whereas the semantic models did not, reflecting that the LSTM lexical prediction after the verb mainly determined likely syntactic frames, assigning high probability values to a number of function words. Second, neither a subject noun nor a verb semantics showed strong correlations at the point when they are revealed but only the verb semantics model showed a strong peak at the point of a function word in the complement phrase in conjunction with other semantic constraint models. This means that the similarity pattern of the semantics of verbs was strongly related to that of lexical prediction on the complement content word in LSTM, implying the importance of a verb in determining the semantics of its complement. This finding is particularly informative because it suggests that the prediction on the complement noun is strongly determined by the verb semantics, showing higher correlation than the full-context semantic constraint model in orange (see below for further discussion).

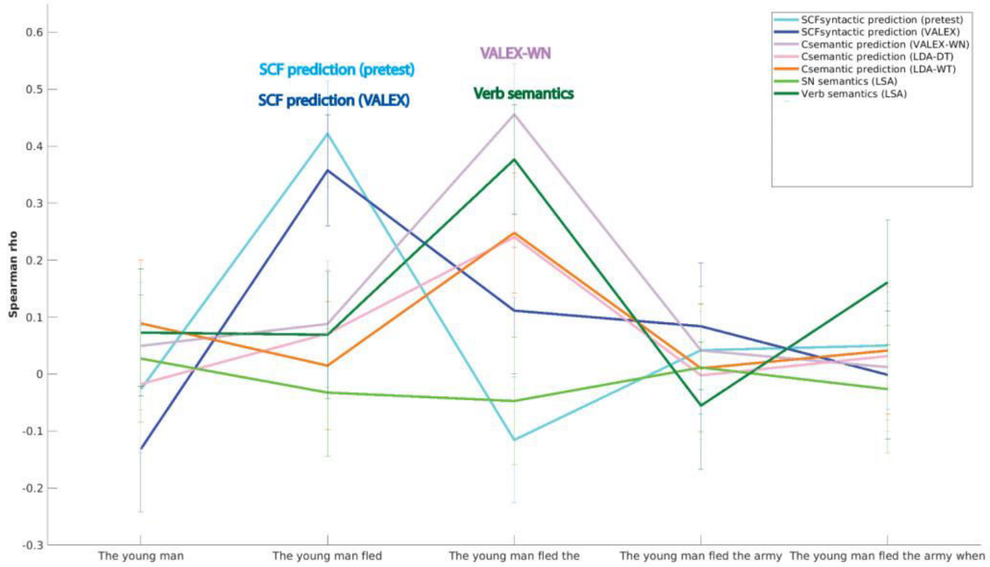


Figure 4-3: A correlation plot of the softmax output layer. Other annotation details are same as in Figure 4-1.

4.3.3. Summary and discussion: Comparing LSTM with human speech processing

The above analyses show the nature of information processing in the LSTM internal layers to generate an output prediction regarding an upcoming word. Especially, the pattern decoding approach revealed the dynamic transition of various linguistic information and incremental computations at different points in a sentence. In summary, there are two important points that must be highlighted from these results: 1) syntactic constraint fits are specific to the verb (generally weaker than semantic constraint fits), consistent with the brain imaging results and 2) the verb semantic effect appeared even at the point of the complement function word along with the semantic constraint fits on its complement.

The first point emphasises the restricted importance of the syntactic aspect of the constraint which is expectedly activated specifically in prior to the specifier of the complement phrase. This is a point where a phrasal structure is constructed by opening up a node that consists of every constituent in the phrase. In hierarchical rule-based accounts, the phrasal constituents can be recursively analyzed into a number of specifier-head configurations (or mini-phrases) whose maximal projection is eventually merged with the specifier that opened up the phrase through a number of bottom-up projections. Hence, it is meaningful to observe the effects of syntactic constraint on the complement specifically at the point of a verb as it implies that the syntactic understanding of a sentence is initiated by activating the structures that frequently co-occur with the context. This is highly consistent with the spatiotemporal patterns of neural activity showing significant effects of the syntactic constraint from 170ms to 500ms after the verb onset in the left fronto-temporal language network (Tyler & Marslen-Wilson, 2008). Further, the syntactic constraint effects were more strongly correlated with the softmax prediction at the verb than the internal layers possibly because the semantic constraints are the primary source of predicting various content words in the vocabulary whereas the syntactic constraints are mainly useful in predicting function words. As a result, the syntactic constraint fits are very specific to the verb and are generally weaker than semantic fits for generating lexical prediction.

In contrast with the full-context semantic constraint effects in humans as early as the subject noun described in Figure 3-5, these effects were only observed after the verb. Further, a strong fit of the verb semantics model was observed at the point of generating the lexical prediction on the following content word in the complement. These results converge to a claim that the LSTM network uses the verb as a primary source of constraints on the complement semantics. Supporting this claim, it was also shown that the clustering patterns of different contexts (subject noun phrases) represented in the LSTM internal layers drastically change based on the verb. Although the recurring theme is saved in the network, enabling it to refer back to distant words in the context (see example LSTM sentences in the section 4.3.1), the actual prediction on the complement semantics is largely determined by the preceding verb. This may reflect the limitation of a predictive machine as a descriptive model since the ultimate goal of human speech comprehension is to understand a message that a speaker wants to convey, not to make an accurate prediction of an upcoming word. Hence, the semantic constraints on the complement are constructed as soon as the theme appears in a sentence in humans, unlike an incremental predictive machine that utilizes the semantic constraints strictly at the point of prediction (i.e. just before the target appears).

4.4. Neurocognitive hypotheses: characterization of the neural response patterns using the LSTM layers

Following on from the section 4.3 exploring the information represented in different layers of the LSTM network at different points in a sentence, this section further investigates how well information processing in the LSTM network captures the spatiotemporal dynamics of neural activity. The representational properties of different LSTM layers were tested against the source-localized EMEG data using RSA to illuminate the similarities and differences in sentence processing between a predictive machine at different layers and human brain. All statistical analysis procedures were exactly same as described in the section 3.2 in order to prevent the variation in the results due to methodological differences.

To address these questions, the three sets of models from different LSTM layers were constructed and tested at different points in a sentence based on hypotheses. First, the two LSTM hidden layers (HL0 and HL1) and the LSTM output layer (softmax) at the point of the subject noun (e.g. "*The experienced walker*") were used to construct model RDMs to characterize the spatiotemporal patterns of neural responses before the onset of a verb (Epoch V at AP-V in Figure 4-4). Further, it was shown from my previous analysis (Figure 3-5) that the contextual constraint on the complement semantics emerges as soon as the theme of the subject NP is revealed. Hence, it was hypothesized that the softmax prediction on the complement semantics could have an effect as early as the subject noun whereas the prediction on the complement syntax would have an effect around 200ms after the verb-onset, which is when the verb-alone and full-context syntactic constraint effects emerged in Chapter 3 (Figure 3-6). From these hypotheses, analyses in this epoch included the models constructed from the output layers at the point of a verb (e.g. “The experienced walker chose”) capturing the constraint on the complement syntax and at the point of the direct object determiner (e.g. “*The experienced walker chose the*”) constraining the complement semantics. With these five model RDMs, I aimed to evaluate the evidence that speech comprehension in human is incrementally predictive, yet is not limited to the immediately adjacent input as in this neural network model and extends beyond adjacent linguistic units through cognitive operations such as utilization of event representation and pragmatics which necessitates an early utilization of semantic (or thematic) constraints based on the SNP on the complement.

Next, these softmax predictions at the point of a verb and a determiner “*the*” were also tested at the epoch aligned to the onset of the complement phrase (Epoch CFW at AP-CFW in Figure 4-4), the point where the main verb in a sentence is revealed. Hence, this is the point where these models should fit the pattern of neural responses if human predictive processing is truly word-by-word like the LSTM network. The patterns of activation in the internal layers at different points in a sentence were tested in their associated epoch in order to investigate how similarly a spoken sentence is processed in human brain compared to the LSTM network. Therefore, at the epoch aligned to the onset of a content word in the complement (Epoch CN1 at AP-CN in Figure 4-4), the three sets of models were tested based on two internal layers and an output layer at this point. To be consistent with the analyses in Chapter 3, I only included the direct object trials from this point where the sentence stimuli begin to syntactically vary. The last epoch was aligned to the onset of the content word (Epoch CN2 at AP-CN in Figure 4-4) in order to test the sentence processing in human brain against the trained network when the three key thematic units are identified, which are the main pillars of constructing a mental model and event representation.

4.4.1. Implications of the section 4.3 on neurocognitive hypotheses

If the way that a sentence is processed in the network is similar to the way that it is processed in the brain, we will likely observe similar results as in the section 3.6. However, since it is already shown and discussed that the effects of constraints in the brain are not specific to the point just before the target word is heard, I expect to observe relatively weaker effects of the

LSTM lexical constraint on the complement around the verb-onset as in the full-context semantic constraint model (Figure 3-5). I expect these LSTM effects to be weaker in general because the LSTM prediction on the complement was shown to be more strongly influenced by the preceding verb and its lexical properties, indicating a fundamental difference in generating constraints. Further, I expect a different pattern of results between HL0 and HL1 since the information encoded in HL0 reflects the general lexical semantics of an input word whereas HL1 captures the multi-level linguistic properties of constraints (see Figure 4-1 and 4-2). Hence, HL0 may characterize the pattern of neural responses associated with lexical processing of an input similar to Figure 3-9 whereas HL1 may lead to a similar pattern of results as shown in Figure 3-5. Lastly, following on from 4.3, my prediction is that these LSTM models fit strongly in the right fronto-temporal areas involved in semantic computations of prediction and integration (Jung-Beeman, 2005), except for the LSTM softmax prediction model at the point of a verb which will involve the left fronto-temporal regions involved in activating the lexico-syntactic properties and constructing a syntactic structure (Tyler et al., 2013).

4.4.2 Epochs and analysis

Following on from the analysis in Chapter 3 using behavioural and corpus-based computational models, this chapter thoroughly explores the various incremental computations through a number of processing layers in the network and quantifies the explanatory values of the recurrent network on the spatiotemporal dynamics of neural activity. Since I analyzed the same dataset with the same analysis pipeline (but with different models), all methodological details are as described in 1.6 and 3.2. Further, epochs were generated to investigate the dynamic changes in neural computations through incrementally unfolding words and their associations with the network's computations. Hence, starting from a subject noun to the complement noun, the spatiotemporal patterns of neural activity were characterized using the patterns of the network's internal state at the particular epoch. In order to directly compare the performance of the network in modelling brain activity with that of behavioural and corpus-based models, the network's prediction at the output layer was also used to characterize the patterns of neural activity. Following on from the results that the complement is semantically constrained as early as the subject noun; the network's prediction on the complement was tested from the epoch that includes the subject noun. See Figure 4-4 for details.

An additional analysis was carried out to investigate utilization of the constraint captured by the LSTM prediction in relation to the processing of the target word. To be consistent with the previous analysis in Chapter 3, I calculated the distance between the LSTM prediction on the complement noun and the actual complement noun (i.e. surprisal) to quantify the difficulty of processing and tested this model of surprisal against the neural activity aligned to the onset of the complement noun. Further, activation of neurons in the first internal layer at the point of the complement noun was also used to characterize the pattern of neural responses at this epoch (see Figure 4-4). In conjunction, the temporal profile of correlation time-courses between each of these models with neural activity illuminates the spatiotemporal dynamics from utilizing the predictive information to processing the target word with respect to the context.

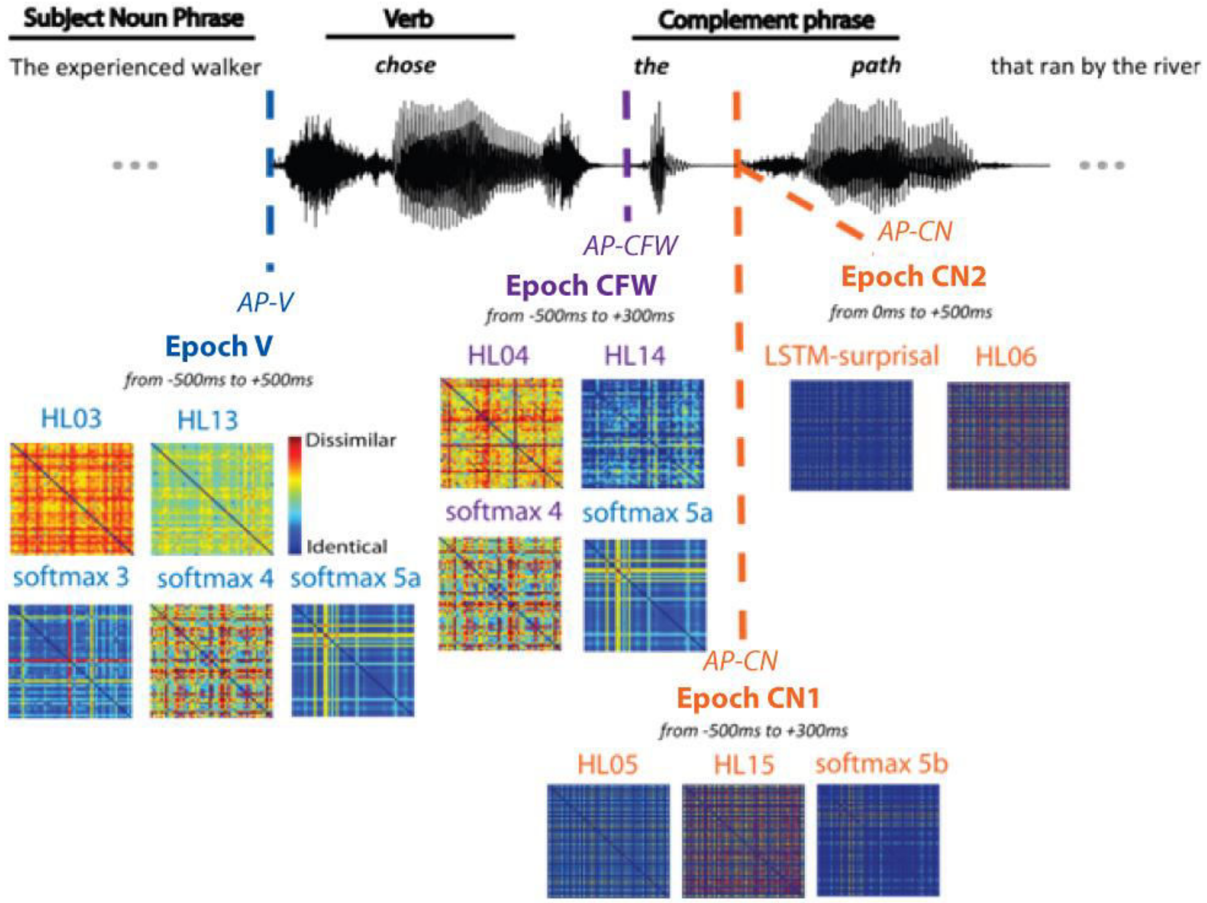


Figure 4-4: Overview of the epochs tested in the experiment in relation to the LSTM models associated with different points in a sentence derived from each of the three layers in the LSTM. The four epochs were each defined relative to one of the three alignment points (AP), with AP-V aligned to the main verb onset in blue ("chose"), AP-CFW aligned to the

complement phrase function word onset in purple ("the") and AP-CN aligned to the complement phrase content word onset in orange ("path"). Each AP is marked on the waveform as a vertical broken line. Epoch CN1 and CN2 were both aligned to AP-CN at different time windows written in Italic. The visualized model RDMs depict the epoch at which they are tested.

Abbreviations: HL stands for "hidden layer" and a number appearing next to it describes each of the two different hidden layers (0 = a first hidden layer receiving an input and projecting to a second hidden layer; 1 = the second hidden layer projecting to the output layer). Then, another number that follows shows how many words are contained in the context. For example, "HL03" represents the state of the first hidden layer after receiving a subject noun (or a third word in a sentence) as an input. Softmax stands for the output prediction and, since there is only one output layer, the number appearing next to it describes the number of words in the context; for example, softmax 3 represents the network's output prediction at the point of a third word in a sentence. Lastly, softmax 5 has two versions: one (softmax 5a) including all trials and the other (softmax 5b) including direct object trials only (recall that there are 100 direct object sentences out of 200 sentences with varying complement structures).

4.5. Results

The analysis at 4.3 revealed the linguistic information encoded in each of the layers in the network. Using RSA, the pattern of information encoded in each of the layers was used to characterize the spatiotemporal dynamics of neural activity across the ROIs tested in Chapter 3. Using the model of the first internal layer (HL0), the lexical-semantic properties of an input represented in the context of the preceding fragment was tested; for example, HL03 was used to test the semantic properties of a subject noun in the subject NP and HL04 was used to test the semantic properties of a verb in the context of the preceding subject NP. Next, using the model of the second internal layer (HL1), the predictive state of the system that strongly represents the constraints on the upcoming words at the abstracted, compact dimensions was tested on the source-localized EMEG data; for example, HL13 was used to test the predictive state when a subject noun in the subject NP is heard and HL14 was used to test the predictive state when a verb in the context of the subject NP is heard. Lastly, using the model of the output layer (softmax), the actual word-level prediction defined over nearly 800,000 types was used to capture the constraints activated in different brain regions; for example, softmax3 was used to test the constraint after the subject NP is heard and softmax4 was used to test the constraint after the subject NP and the verb are heard.

By testing these models derived from the LSTM layers at each word from a subject noun to a complement noun, I aimed to elucidate the influence of each incrementally unfolding word on predictive state of the brain and the way it constrains the upcoming word. Further, the results from these network models are compared with respect to the results from behavioural and corpus-based models tested in Chapter 3 to highlight the potential difference in the way that each word is constrained and processed between the network and the brain. The epochs at which each LSTM model is tested is described in Figure 4-4. Again, all other methodological details (data pre-processing, statistical tests and multiple comparisons correction and ROIs used in the analysis etc.) were exactly same as described in Chapter 3; the only difference was the models and the epochs at which the models were tested.

4.5.1. (i) Subject Noun and Verb (Epoch V):

In order to determine the relationship between the network's incremental computations and the brain data at different key points throughout the sentence, I correlated the spatiotemporal patterns of neural activity with the two hidden and one output layers (for more discussion about linguistic information encoded in each of these layers, see section 4.3). As expected, all three layers showed completely different patterns of model-fits, and all of the effects were in the right hemisphere, consistent with the semantic effects of the pre-test and corpus-based models in Chapter 3.

In section 4.3, I showed that information processing in the first hidden layer (HL0) involves activating the semantics of an input word. In line with this result, significant effects of HL0 were observed mainly in the right posterior temporal cortex (see Figure 4-5). The effect initially emerged in RAG around 400ms before the verb-onset, followed by even stronger fits of the posterior temporal regions, extending anteriorly to RaSTG/MTG and rostrally to RBA44 (RAG: p=.004 peaking at -375ms; RBA44: p=.013 at -330ms; RaSTG: p=.027 at -290ms; RaMTG: p=.018 at -170ms; RpSTG: p=.002 at -130ms; RMTG: p=.022 at -120ms; RHG: $p < 0.001$ at -100ms). This temporal transitioning of the model-fits among different regions possibly reflects the usefulness of the input semantics in early construction (RAG – RBA44) and computational refinement of the constraints (RaSTG/MTG) for lexical prediction (RpSTG/MTG and RHG). Also, it is important to note that a small peak appears again in the later stage around 300ms after the verb-onset in posterior temporal areas (only significant in RHG: p=.012 at 290ms but other regions including RpSTG/MTG and RaSTG showed similar patterns as well). This is likely to reflect a thematic recurrence for constructing the mental scenario or message-level representation from which the constraints are generated.

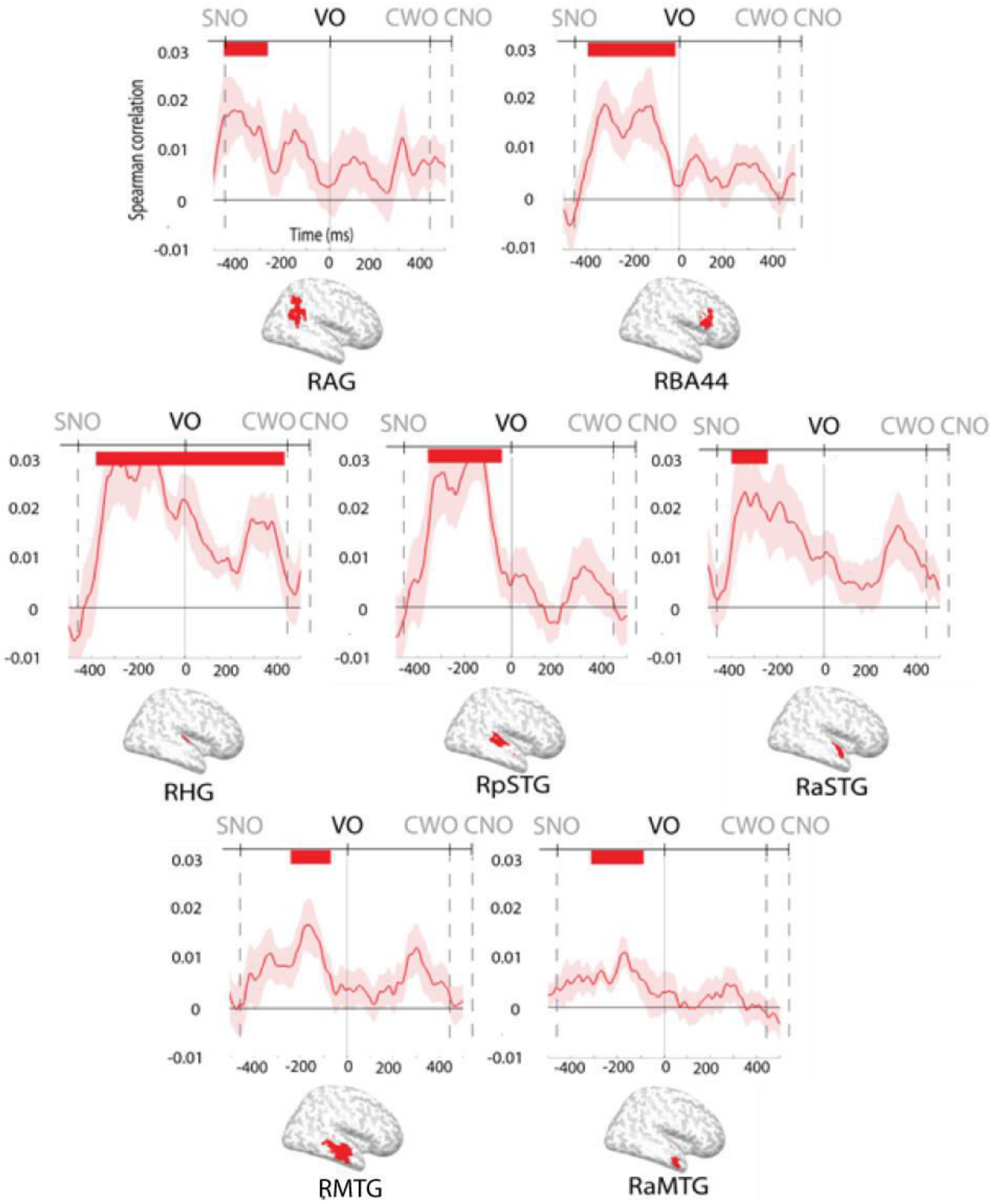

SNO
VO
CWO CNO
0.03
0.02
0.01
0
-0.01
Spearman correlation
Time (ms)
-400
-200
0
200
400
RAG
RBA44
RHG
RpSTG
RaSTG
RMTG
RaMTG

Figure 4-5: ROIs with significant RSA model fits in Epoch V for HL03 (the first internal layer given three words in the context (i.e. subject NP). The Spearman correlation time-courses for HL03 model fit for the 7 significant ROIs (R Heschl's gyrus (RHG); R posterior Middle Temporal Gyrus (RMTG); R anterior MTG (RaMTG) R posterior Superior Temporal Gyrus (RpSTG); R anterior STG(RaSTG); R BA 44). The time periods of significant model fit are indicated by red bars across the top of each ROI plot. These values are corrected for multiple comparisons across time using threshold free clustering enhancement (TFCE). The VO alignment point is marked by the black horizontal line in each figure whereas the dotted line reflects the estimated onset of a particular word from the average word durations across trials (indicated by the abbreviation above the plot in grey). The shaded ribbon on either side of the red correlation line indicates standard error across subjects. SNO = subject noun onset, VO = verb onset, CWO = complement function word onset and CNO = complement content word onset.

I showed in section 4.3 that the second hidden layer (HL1) constructs multi-level constraints on the upcoming word based on the projected semantic information of the preceding context (word) from the HL0. Interestingly, only two regions showed significant correlations with the state of HL1 (Figure 4-6). First peaks appear around 350ms before the verb-onset in RMTG and R-BA44 (RMTG: p=.041 at -350ms; R-BA44: p=.038 at -330ms). The peak in R-BA44 lasted until 250ms before the verb onset and reappeared 50ms later (R-BA44: p=.049 at -150ms) whereas the peak in RMTG disappeared soon after and reappeared again around 200ms before the onset which only lasted for about 40ms (RMTG: p=.04 at -185ms). It is worth noting that the patterns of these model-fits have similar temporal dynamics in the same regions as the earlier first-layer computations associated with HL0 except that the first peak in RMTG does not reach significance in the HL0 fit.

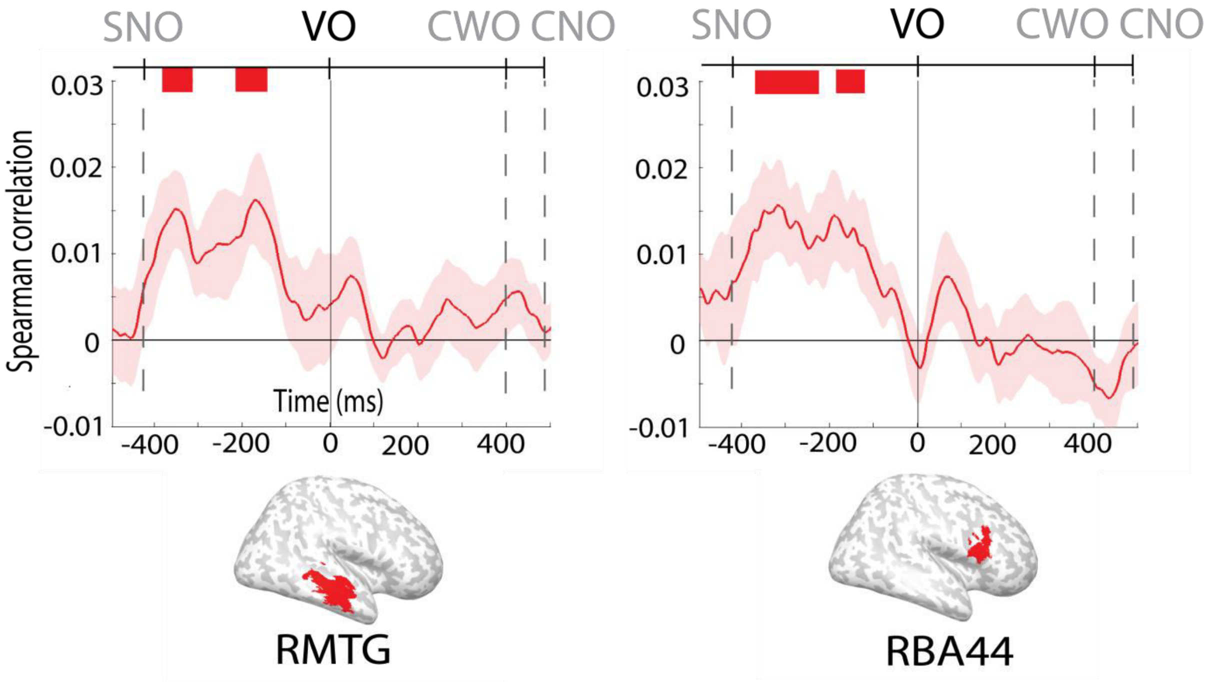


Figure 4-6: ROIs with significant RSA model fits in Epoch V for HL13 (the second internal layer given three words in the context (i.e. subject NP). The Spearman correlation time-course for the HL13 model fit in RMTG and R-BA44. Red bars indicate TFCE-based significant model fit. The peak fit in each region is as follows: R-BA44 at -330ms and RMTG at -185ms. Details and legend are as in Figure 4-5.

Next, I compared the pattern of similarity between the output layer and the brain activity aligned to the onset of the verb (see Figure 4-7). However, following on from the finding that brain constrains the semantics of the upcoming complement as early as the subject noun, I further tested the neural activity against the response patterns in the output layer at every incremental point from the subject noun to the complement function word. First, the output prediction regarding the upcoming word after the subject noun phrase (Softmax3) showed significant model fits in two different regions including RaSTG and RAG (RaSTG: p=.015 at -340ms; RAG: p=.031 at -145ms). Interestingly, none of the regions involved in computing the constraints at HL1 (RMTG and R-BA44) showed significant correlations with the projected output although all of these four regions including RaSTG and RAG were involved in early semantic activation in HL0.

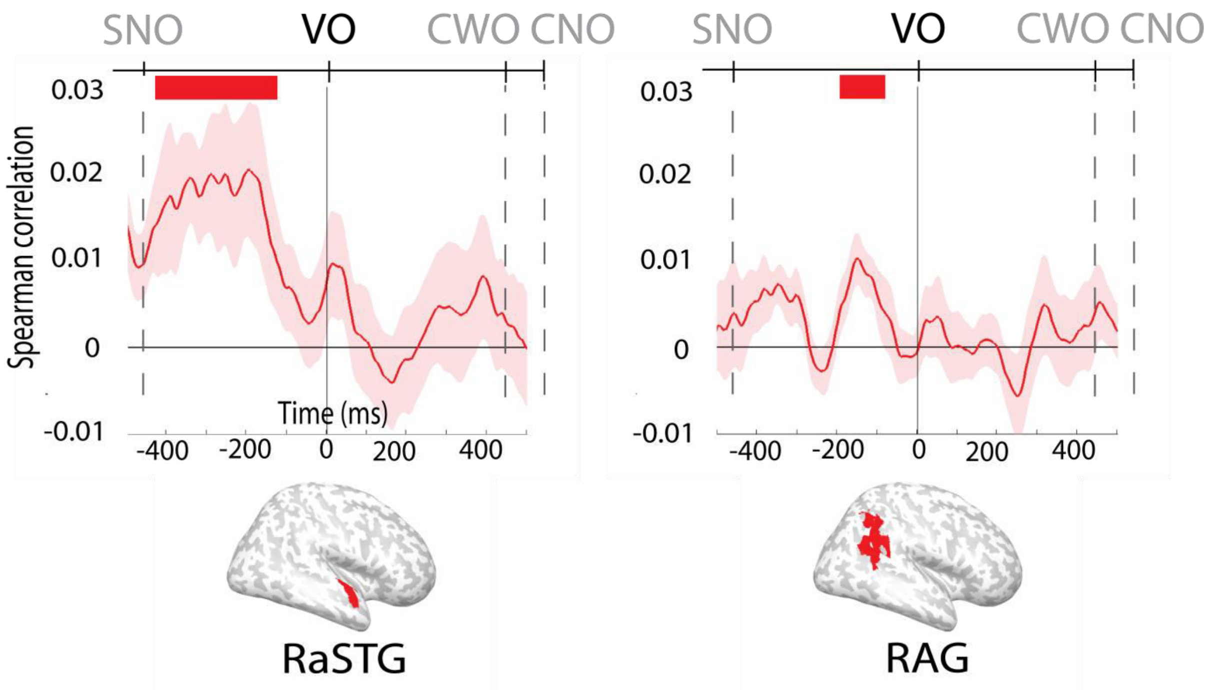


Figure 4-7: ROIs with significant RSA model fits in Epoch V for Softmax3 (the output softmax layer given three words in the context (i.e. subject NP)). The Spearman correlation time-course for the Softmax3 model fit in RaSTG and RAG. The peak fit in each region is as follows: RaSTG at -340ms and RAG at -145ms. Details and legend are as in Figure 4-5.

As expected, the response patterns in the output prediction of the complement noun (Softmax5) showed significant correlations at similar time in similar regions. I constructed the models in two different ways, first based purely on the softmax distribution in the output layer just as the Softmax3 model and second based on a semantic blend of 50 most likely candidates predicted by the softmax output using the topic-word vectors (to minimize the methodological differences between the behavioural and the LSTM constraint for comparisons). A semantic blend effect first emerged in RHG as early as 320ms before the verb-onset lasting for about 200ms (this effect showed two peaks; RHG: p=.038 at -300ms and p=.026 at -200ms). Following on from this early semantic effect, more specific constraints at the lexical level appeared in the right anterior temporal regions in RaITG and RTP transitioning into R-BA44 around the verb-onset (RaITG: p=.006 at -130ms; RTP: p=.004 at -80ms; R-BA44: p=.038 at 90ms). This pattern of results suggests that the network's constraint on the complement noun semantics captures the early activity in RHG, possibly when the constraint starts to be constructed. However, as listeners hear hundreds of milliseconds into the subject noun, the constraint becomes fine-grained and specific, modelling the activity pattern in the anterior temporal areas (RaITG and RTP) until 20ms before the verb-onset and in the RBA44 soon after the verb-onset (Figure 4-8). These patterns of results are generally consistent with the behavioural model (i.e. full-context constraint model in Chapter 3) showing effects centred in the anterior temporal regions for computation, and then transitioning into the inferior frontal regions possibly for selection.

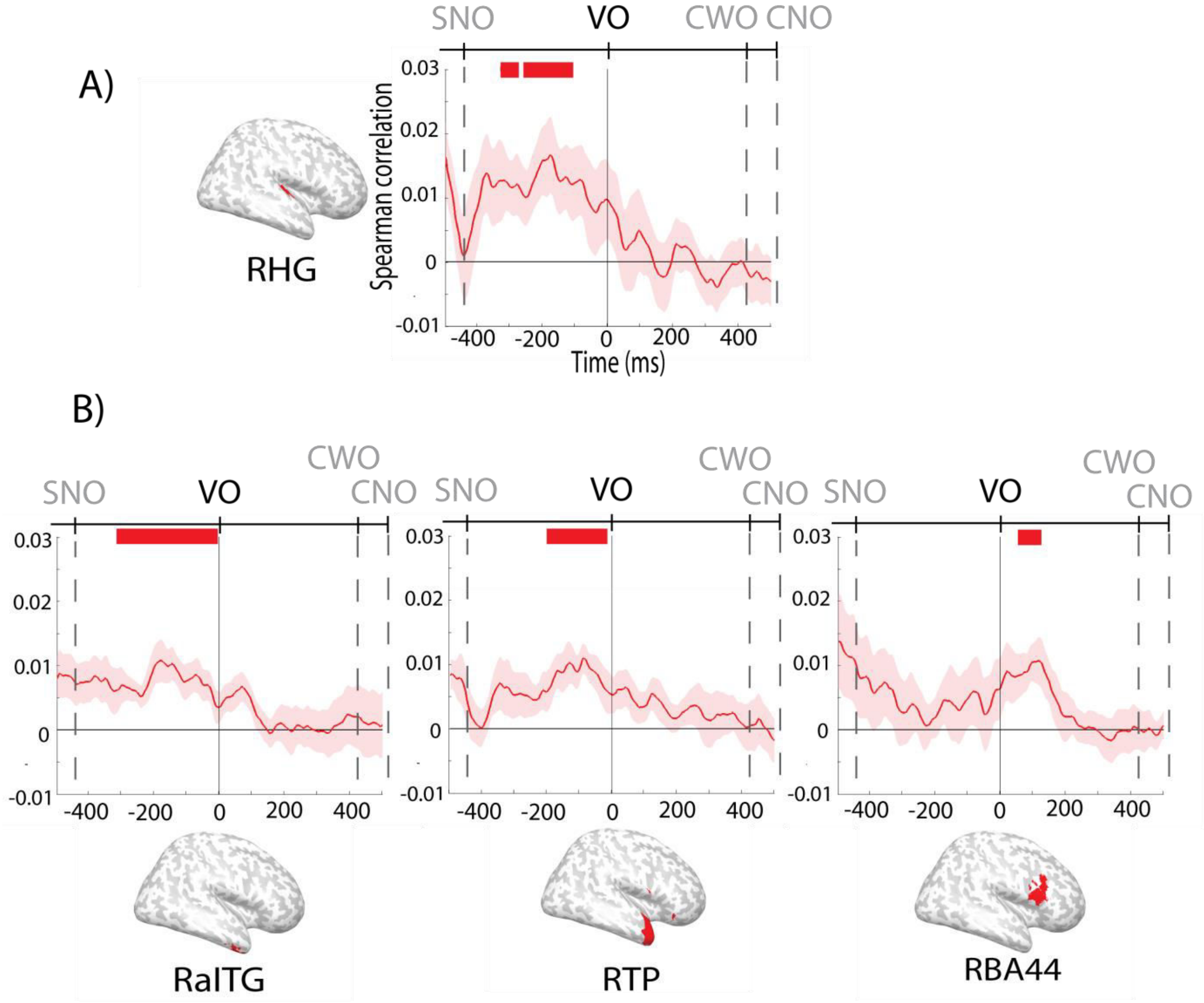


Figure 4-8: ROIs with significant RSA model fits in Epoch V for Softmax5 (the output softmax layer given five words in the context (i.e. subject NP + verb + 'the')). Again, these predictive models on the complement content word are tested at this epoch aligned to VO as it has been shown in Chapter 3 that the complement semantics starts to be constrained while a subject noun is being heard. The Spearman correlation time-course for the blended Softmax5 model fit in RHG (panel A) and the pure Softmax5 in RaITG, RTP and RBA44 (panel B). The peak

fit in each region is as follows: RHG at -200ms (panel A); RaITG at -130ms, RTP at -80ms and RBA44 at 90ms (panel B). Details and legend are as in Figure 4-5.

4.5.2. (ii) Complement onset (Epoch CFW):

Unlike Softmax3 and Softmax5, the model of LSTM prediction based on 4 words in the context (i.e. SNP + verb) denoted as Softmax4 did not show any significant effects. None of the brain regions showed similar patterns of activity regardless of the epoch that the data were aligned to (Epoch V and Epoch CFW in Figure 4-4). Consistently, neither HL04 nor HL14 showed significant correlation in any of the ROIs aligned to the complement onset (Epoch CFW in Figure 4-4). This may possibly reflect that the predictive processing in humans does not occur for every word in a sentence but occurs only for the content words that have thematic significance in terms of constructing the message-level representation. Given that the Softmax5 model showed significant correlation time-courses in the right anterior temporal and BA44 regions, the absence of correlation for the Softmax4, HL04 and HL14 models may possibly indicate that the network's computation of syntactic constraint (or, more precisely, lexical constraint on a specifier of a phrase that often determines the phrasal structure in a sentence) is not very similar to the way that humans constrain the syntactic structure of a complement.

4.5.3. (iii) Complement noun onset (Epoch CN1):

Following on from the earlier fit of the LSTM prediction model of the complement noun (i.e. Softmax5) aligned to the verb-onset, it reappeared in RaSTG from -160ms before the onset of the complement noun lasting for 140ms (RaSTG: p=.024 at -45ms; see Figure 4-9). However, the first (HL05) and second internal layers (HL15) were correlated with none of the ROIs in this study. This suggests that the neural computation involved in generating constraint is different from how the network generates word-by-word constraint which possibly reflects that the predictive processing in human brain does not occur at every point in a sentence specifically at the lexical level.

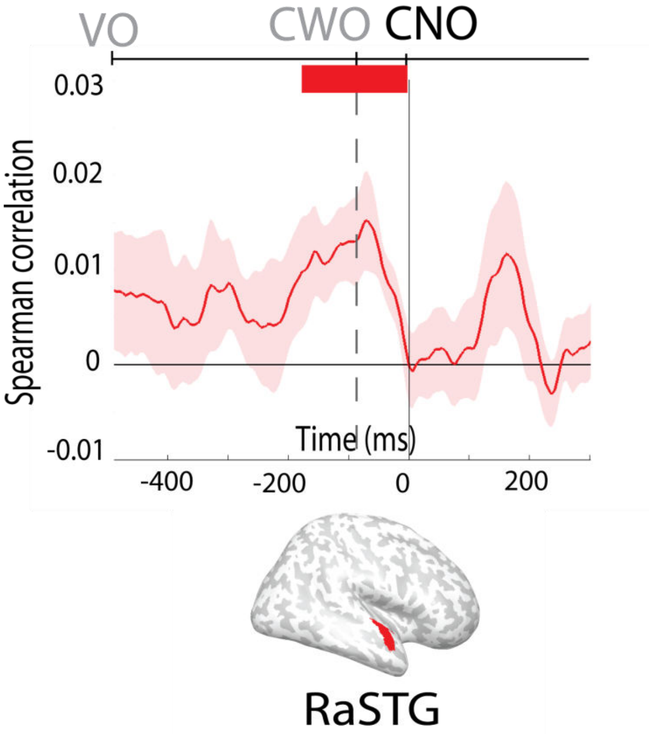


Figure 4-9: ROIs with significant RSA model fits in Epoch CN1 for Softmax5 (the output softmax layer given five words in the context (i.e. subject NP + verb + 'the')). The Spearman correlation time-course for the Softmax5 model fit in RaSTG (panel A). The peak fit in this region was at -45ms. Details and legend are as in Figure 4-5.

4.5.4 (iv) LSTM models vs. Behavioural and corpus-based models

The LSTM network used in this study is an incremental model of language processing trained to predict the upcoming word as accurately as possible. Unlike rule-based computational models, it does not have an explicit knowledge of syntax. Unlike corpus-based models, it captures the crucial aspect of incrementality in sentence processing via recurrent projection between the internal layers. In this section, I describe how well such LSTM model with a simple architecture and a straightforward representation capture the spatiotemporal dynamics of neural activity in comparison with the behavioural and corpus-based models.

Starting from the early computation occurring in the first internal layer, this model captured the activity patterns in a number of ROIs centred on the right posterior temporal lobe (See Figure 4-5). These patterns of results are consistent with the lexical semantic activation of the subject noun, modelled by the DM vector (Baroni & Lenci, 2010) in Chapter 2. This corpus-driven semantic vector showed significant correlations in a number of right temporal regions mainly in the superior and posterior areas. Not surprisingly, four different ROIs in these areas consistently showed significant correlation with highly similar temporal profile. These ROIs include RHG, RpSTG, RaSTG and RMTG which are the regions consisting of the processing stream of the lexical semantic information (Hickok & Poeppel, 2007). Further, other adjacent temporal regions including RaITG and RpMTG showed non-significant yet very similar correlation time-courses. In conjunction, these results suggest that the network is capable of capturing the lexical-semantic processing in the right posterior temporal regions on the incrementally unfolding input in its first internal layer. However, given that HL04 and HL05 did not show any significant effect, it requires further research to investigate how well the network's first internal layer captures the neural processing of an input with varying linguistic properties (especially, the syntactically meaningful function words).

Next, to evaluate the network's output prediction against the brain data and to compare its performance in relation to the behavioural (full-context) constraint on the complement semantics, I took the exact same procedure to construct the full-context semantic constraint model (see the section 2.5.2.3) to generate a new semantic constraint model based on the LSTM prediction with minimal methodological difference. The only difference between these two models was that one was based on human prediction whereas the other was based on LSTM prediction (see Figure 4-11 for comparisons). As written above, this LSTM model was significantly correlated only with RHG as early as 320ms before the verb onset which peaked around 200ms before. Unlike the full-context semantic constraint model, the effect did not transition into the other brain regions and the significant model-fit in this region drastically declined around the verb-onset. Instead, other anterior temporal regions (RaITG and RTP) and RBA44 showed significant correlation with the fine-grained LSTM prediction model based on the softmax output distribution. These results imply that the network is capable of explaining the early construction of the semantic constraint in RHG although the later fits require a more fine-grained distribution than a blended vector across top 50 most likely candidates (see 4.6.2 in discussion below)

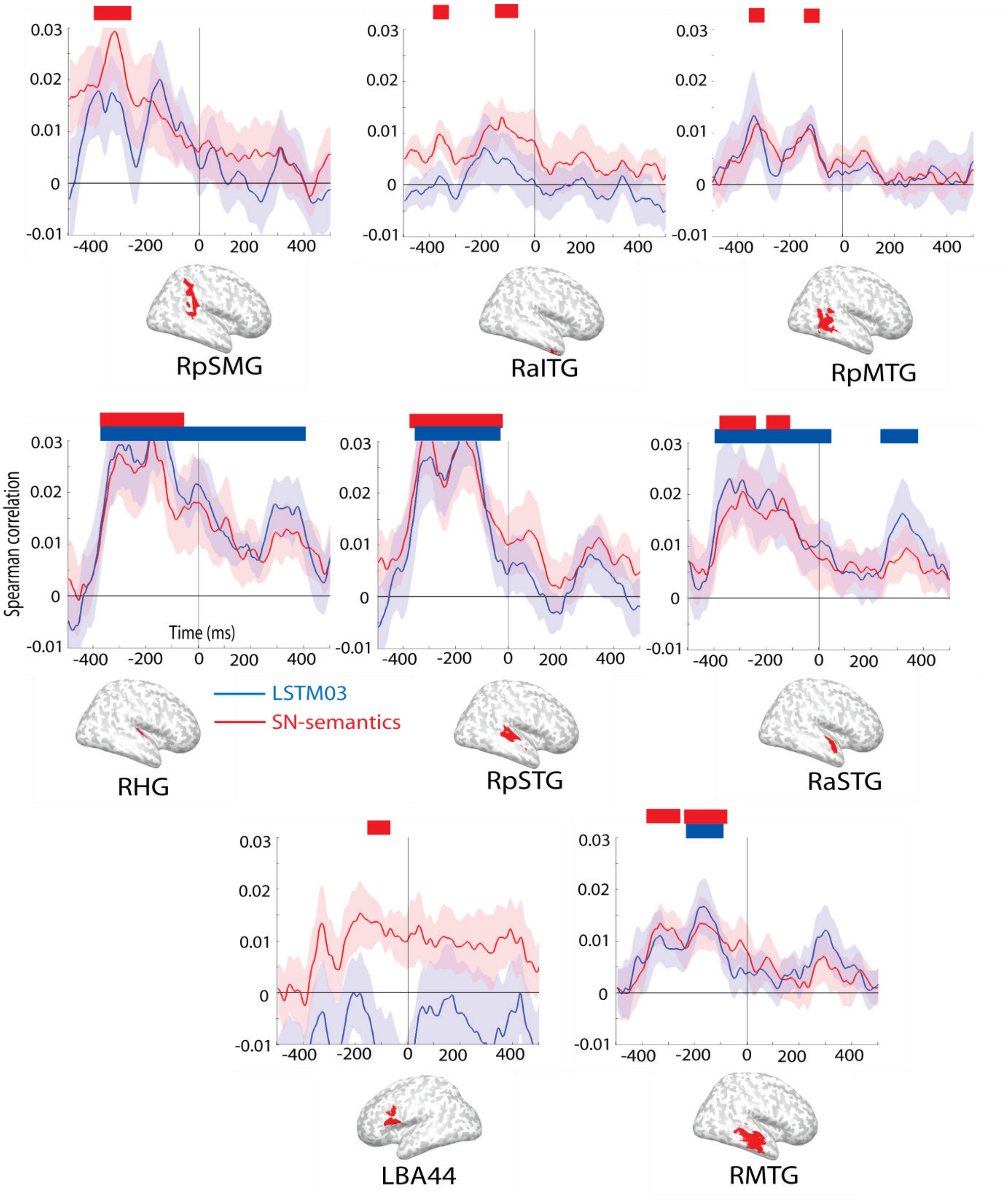


Figure 4-10: ROIs with significant RSA model fits in Epoch CN1 aligned to the verb onset for the lexical-semantics based on the DM vectors (Baroni & Lenci, 2010) compared to LSTM03 (the first internal layer given three words in the context (i.e. subject NP)). They are denoted in red and blue respectively as “SN-semantics” and “LSTM03”.

The Spearman correlation time-course for the LSTM model fits in blue and for the corpus-based model fits in red. Each plot shows a contrast between these fits in each of the eight significant ROIs. Details and legend are as in Figure 4-5.

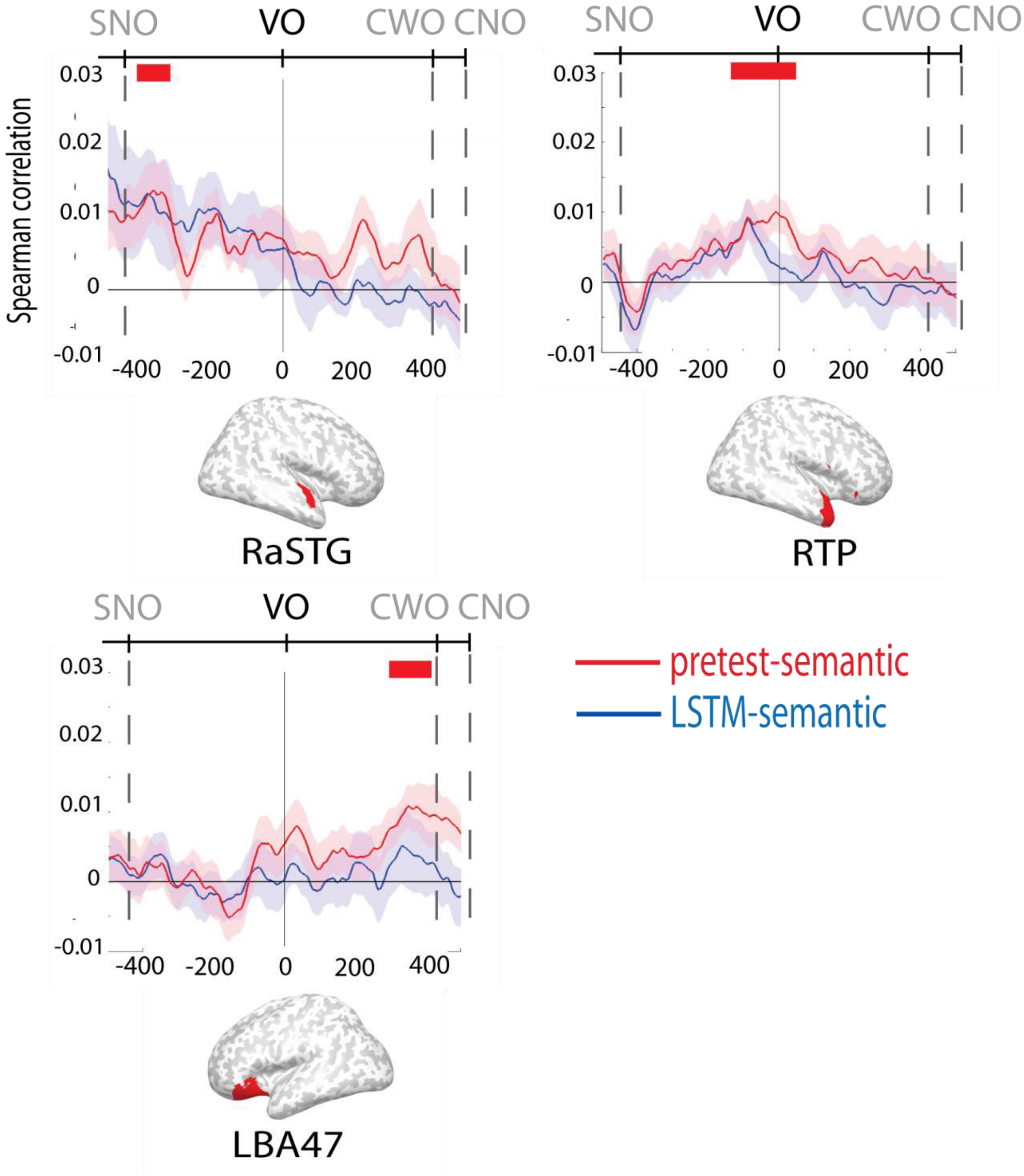


Figure 4-11: ROIs with significant RSA model fits in Epoch CN1 for the full-context semantic constraint based on the human behavioural response in relation to blended Softmax5 (the output softmax layer given five words in the context (i.e. subject NP + verb + 'the'). They are denoted in red and blue respectively as "pretest-semantic" and "LSTM semantics" (the only difference between these models was the use of behavioural vs. LSTM prediction). The

Spearman correlation time-course for the LSTM model fits in blue and for the pretest-semantic model fits in red. Each plot shows a contrast between the pretest fit and the LSTM fit in 3 different ROIs that showed significant effects in the previous study. Details and legend are as in Figure 4-5.

4.5.5. (v) Additional analysis (Epoch CN2):

Lastly, in order to explore the similarities and differences between the brain's and network's utilization of constraint in relation to processing the complement noun, I compared the network's prediction on the complement noun and its first internal state (which was shown to reflect the semantics of the context and the input word more than the constraint on the upcoming units; see 3.3) with the neural activity aligned to the onset of the complement noun. The results showed a significant surprisal model fit emerging around 100ms and peaking at 190ms after the noun onset in LpITG (LpITG: p=.034 at 190ms). Soon after these clusters were observed, a number of significant clusters were found in RITG for the internal state model at the complement noun (HL06) which initially emerged around 170ms and peaked at 340ms (see Figure 4-12). This pattern of results indicates utilization of the constraint at the early stage of processing the target (in LpITG) to guide the interpretation of it in the light of the preceding context (in RITG). Note that these effects were found in the bilateral ITG, the regions involved in semantic processing of a sentence (Hickok & Poeppel, 2004; Rodd et al., 2005; Binder et al., 2009) consisting of a ventral processing stream (Hickok & Poeppel, 2007). Also, the LSTM surprisal effect was found in LpITG, a region adjacent to LpMTG where the surprisal effect from a pre-test (behavioural) model was found; in fact, the correlation time-courses of these models were similar in LpMTG but the LSTM version had larger variance, leading to a non-significant result (see Figure 4-13).

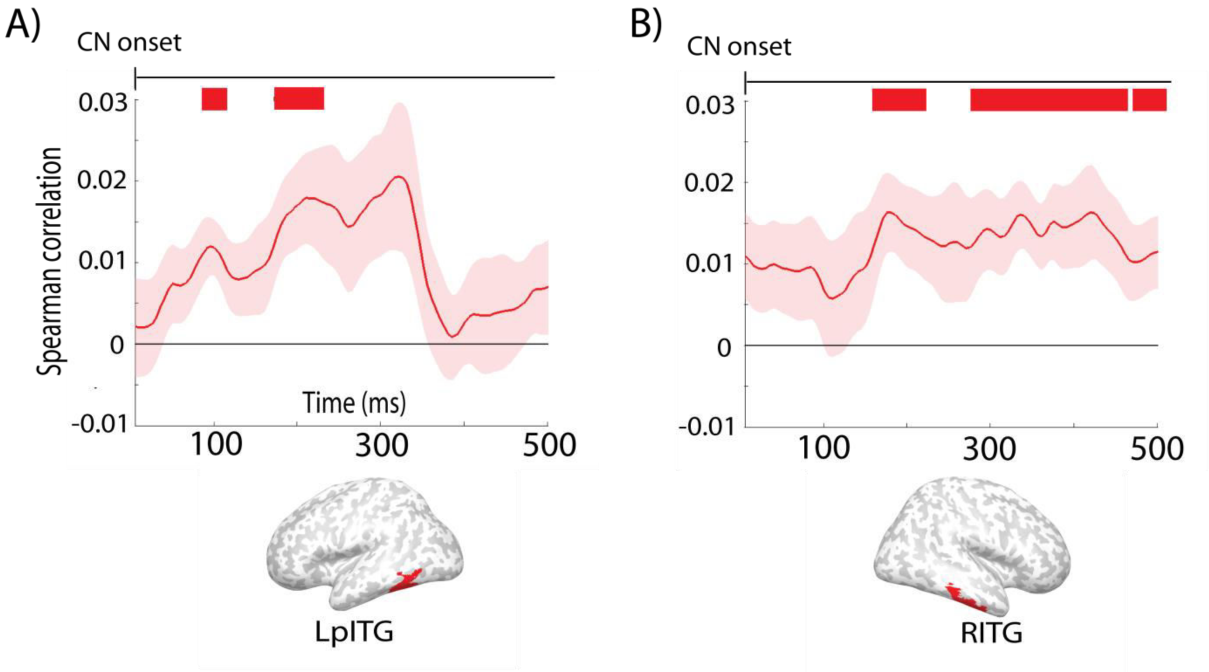


Figure 4-12: ROIs with significant RSA model fits in Epoch CN2 for the LSTM surprisal at CN in left (A) and HL06 (the first internal layer given six words in the context (i.e. subject NP + verb + 'the' + CN)) in right (B). The Spearman correlation time-course for the LSTM surprisal model fit in LpITG and HL06 model fit in RITG. The peak fit in these regions was at 190ms and 340ms respectively. Details and legend are as in Figure 4-5.

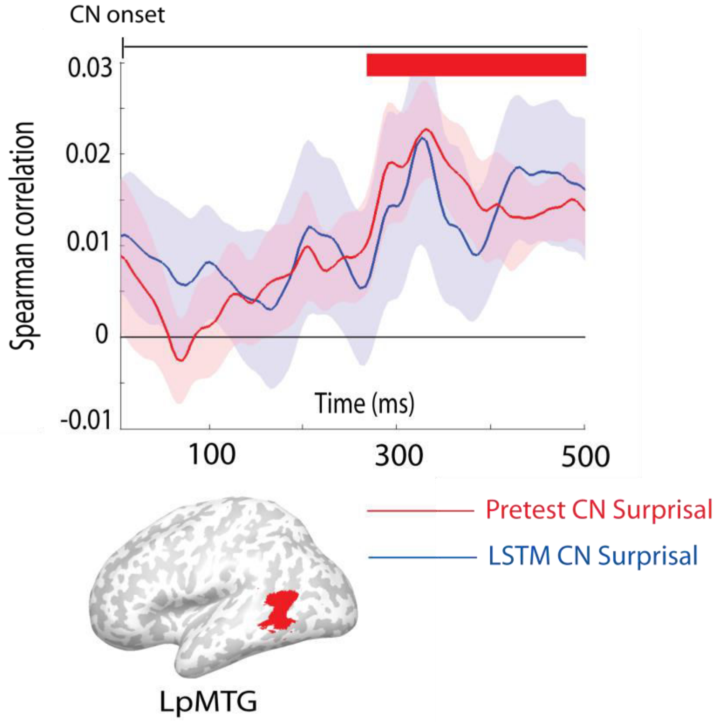


Figure 4-13: ROIs with significant RSA model fits in Epoch CN2 for the behavioural and LSTM surprisal at the complement noun in LpMTG. They are denoted in blue and red respectively as "LSTM CN Surprisal" and "Pretest CN Surprisal". Other details are as in Figure 4-5.

4.5.6. (vi) Summary

Combining the findings across these four adjacent epochs, this study supports the claim that contextual information is essential in predictive processing (Tyler & Marslen-Wilson, 1977; Marslen-Wilson et al., 1993; Nieuwland & Van Berkum, 2006) and suggests that the constraint is constructed early in a sentence which is modified by incrementally unfolding input until the target is heard. Following on from the earlier findings in Chapter 3 with the behavioural and corpus-based models, these findings emphasize that the early neural computation during incremental speech processing can be sufficiently explained by learning statistical regularities in the data even without any syntactic knowledge. Rather, the network highlights the recurrence in predictive processing, allowing the constraint to be computed as early as the subject noun and to be modified through a series of input words for more specific and accurate prediction; in this way, brain is likely to draw the possible scenarios in order to understand the message as quickly and accurately as possible.

In addition to these findings, the absence of model-fits for the internal layers in fourth (verb) and fifth words (complement function word) suggests that the brain is not a strictly word-by-word predictive machine (especially for function words since the Softmax4 model was silent) or that its incremental computations were not well captured by the LSTM network at these points for some other reasons (see 3.6. discussion). Either way, the network's computation throughout the layers characterized the pattern of neural responses only when the input or the target word was a meaningful content word (i.e. a subject noun, a verb and a complement noun) but not when it was a function word that works as a specifier of the complement phrase (having syntactically meaningful information in phrase structure building). Note that there are 8 different complement function words in this experiment so the absence of effects cannot be ascribed to the lack of variability.

4.6. Discussion

The main goal of this study was to address incrementality in human speech processing. An LSTM neural network is a state-of-art connectionist model, designed to investigate how the information represented in different processing layers changes as each word incrementally unfolds over time. Unlike a behavioural or corpus-based model which infers the cognitive state of the brain based on the linguistic properties associated with the output response or massive text corpora, this model is highly flexible in terms of the number of words in the context and is transparent such that the internal state of the system can easily be accessed during incremental processing of language stimuli. In this chapter, I evaluated this connectionist model of language processing against the brain data to probe the nature of neural computation in a predictive framework. Further, I explored to what extent the neural computation can be modelled using the connectionist network that knows nothing about syntax but understands a word in the context of a sentence through recurrent projections.

The dataset used in this study was identical to the previous study in Chapter 3, consisting of a number of source-reconstructed brain regions whose activity was recorded using EMEG while listeners were hearing a set of sentence stimuli. These sentences varied in terms of the probabilistic constraints that a verb generates on its complement. The pattern of brain activity in each of the ROI was characterized using a number of processing layers in the connectionist network (LSTM; Jozefowicz et al., 2016) each of which represents different types of linguistic information. I tested for each of the layers of this LSTM network at every important point in a sentence including a subject noun, a verb, a complementizer (a function word that opens up the phrase structure) and a complement noun at the relevant epoch described in Figure 4-4.

Similar to the current study, a previous MEG study has investigated if it is viable to employ a recurrent neural network (RNN) to decode the time-varying neural activity while participants were reading a story (Wehbe, Vaswani, Knight & Mitchell, 2014). Wehbe et al. (2014) extracted the hidden layer representation, the output word probability and input word embeddings which were used to predict the MEG data for a given word i in their ridge regression analysis using a training set (9-folds). Using these models, they carried out a binary classification task of assigning the label (word i' vs. word ii') to the actual recording i, based on their prediction that more closely matches the MEG recording of the word i in a test set (1-fold). They reported that the classification accuracy was significantly above chance for all models with the hidden layer being most and the output probability being least accurate.

This pattern of results was consistently observed in this study as well, showing relatively extensive and strong fit to the first hidden layer LSTM0 in the RH language network (see Figure 4-5) compared to the second hidden layer LSTM1 (see Figure 4-6) and output softmax probability distribution (see Figure 4-7 and Figure 4-8). From here, it is possible to deduce that the predictive computations in RNN/LSTM capture the neural activity involved in incremental language comprehension, which may start to diverge as the computations are projected towards the output layer. This could be due to the network's output being too fine-grained as its training objective is to make as accurate word-level prediction as possible (see Section 4.6.4 below for more discussion to improve RNN/LSTM in psychological perspective). Regardless, Wehbe et al. (2014) and this study established the validity of using RNN/LSTM as a computational model of human language comprehension in reading and speech.

4.6.1. Activating lexical properties of an input word in a sentence

Understanding a sentence requires an incremental process of activating the lexical semantic information of the input and adapting it in the light of the preceding context. The LSTM network mainly performed this computational process in the first internal layer (see 4.3). This early computation was predominantly observed in right posterior temporal regions, consisting of the ventral stream that works as a lexical-semantic interface (Hickok & Poeppel, 2004). In particular, the four different ROIs including RHG, RpSTG, RaSTG and RMTG showed highly similar correlation time-courses between the corpus-based DM co-occurrence model and the LSTM internal layer model (Figure 4-10). Further, two other ROIs including RaITG and RpMTG were not significantly correlated with this computational process but still showed highly similar correlation time-courses with the Baroni's DM co-occurrence model. All of these effects occurred approximately between -400ms to 0ms; given the average duration of a subject noun is around 450ms, it is clear that this HL03 model captured the lexically specific process in the brain which involves activating a word's lexical-semantics. Further, it might be worth noting that the late cluster appearing around 300ms after the verb onset is generally stronger in the LSTM model fit than in the Baroni's DM model fit, especially in RaSTG which may imply that this LSTM network better captures the recurrent properties of the lexical semantics in the brain for constraining and processing the upcoming input with respect to the distant but important word in the context that constructs a theme in the message.

Three other regions including RAG, RaMTG and RBA44 were only significant for this internal layer model at the subject noun. The effect initially emerged in RAG around 420ms before the verb-onset which likely reflects the early (predictive) interpretation of the subject noun with respect to the preceding modifier. It is well known that bilateral angular gyrus (AG) is involved in representing conceptual semantics of words (Demonet, Chollet, Ramsay et al., 1992; Demonet, Price, Wise & Frackowiak, 1994; Binder et al., 2009; Price, 2012; Kocagoncu et al., 2017). Consistent with this claim, my ongoing analysis recently showed that the contextual semantic representation of the entire subject noun phrase (e.g. "*The experienced walker*") occurs in RAG as early as 500ms before the offset of the phrase which transitions into RpSMG peaking around 330ms before the offset. Similarly, the roles of the bilateral inferior frontal and anterior temporal regions are discussed by Jung-Beeman (2005) suggesting that IFG is involved in selection and control whereas the anterior temporal regions are involved in semantic constraint and integration. In particular, the recent study by Kocagoncu et al (2017) showed that the ease of feature integration of concepts (e.g. integrating the features such as "has stripes", "has four legs" and "eats grass" into a concept "zebra") was significantly correlated with the bilateral AG, RMTG and RIFG (centred on RBA44) around the uniqueness point of a word from a single word listening study. From these results, the early lexical semantic effects in these right posterior parietal regions (RAG, RpSMG) are likely to reflect the contextual semantic representation driven by the preceding modifier whereas the anterior temporal and inferior frontal effects are likely to reflect the integrative and selective processes in order to understand a word (subject noun) in the context of the preceding words (modifier).

Despite this extensive activation in the right hemisphere for lexical semantic processing of a subject noun, such activation was not observed for the subsequent words in different grammatical categories (i.e. a verb and a function word). In fact, there is no semantic information in a function word whereas a verb embodies semantic information which is not as concrete as a noun. From my previous study, it was claimed that a verb is central to constraining the complement syntax whereas it only plays a confirmatory role in constraining its complement semantics. Hence, the absence of model-fits at these points in a sentence is likely because the LSTM network processes these syntactically informative words differently from the brain (see 3.6.4). Consistently, it was suggested that the neural loci of verb processing are different from those of noun processing (Perani, Cappa, Schnur et al., 1999).

4.6.2. Incremental constraint and prediction

The LSTM network constructs a constraint on the upcoming word using the contextual information represented by the first internal layer. As shown in Section 4.3, the second internal layer represents the semantic constraint on the upcoming word (but syntactic constraint to a lesser extent). The regions that were in the similar state to this second internal layer involved RMTG and RBA44 while processing a subject noun. These are the regions discussed above in 4.6.1 which are likely to be involved in constraining the upcoming verb based on the integrated representation of the subject NP (see Kocagoncu et al., 2017). Consistent with the first internal layer, no other significant fits were observed at the point of a verb or a function word for the second internal layer (see discussion above in 4.6.1). Lastly, a direct modelling of lexical constraint from the LSTM output prediction (softmax) showed significant fits in RaSTG and RAG at the point of a subject noun; the brief and relatively late activation of RAG around 140ms before the verb onset may possibly reflect a confirmatory activation of constraint with respect to the contextual semantics. Again, both of these regions showed a correlated activity pattern with the first internal layer.

One of the most interesting findings from these results is that the timings of these fits between the three layers are not very different. Unlike computations in a connectionist network defined by a series of weighted projections with clear non-linear activation functions, neural computations are not easily tractable or mathematically expressible. The best way to infer causality (or a direction of projection) among various computations is to align each computation by the time at which it occurred (no current computation can be projected to the past computation). In fact, this is a basic assumption of various causality analyses such as Granger causality (Ding, Chen & Bressler, 2006; Barnett & Seth, 2014), often used as a measure of directed connectivity in the brain. Hence, if the brain follows the same computational procedure as the LSTM network to compute the constraints, it is expected to observe these effects at least after the model-fit for the first internal layer emerges. However, the earliest effect for each of these layers occurs almost simultaneously around 400ms before the verb-onset. This highlights that neural computations involved in constraining the upcoming input are not necessarily divided into different processing stages that strictly combine the input representation with the previous state of the system (that involves a representation of the preceding context) before making predictions on the upcoming word. As discussed below, this is the main reason why the LSTM network cannot become a descriptive model of human speech processing.

In Section 4.3, it was demonstrated that the network's prediction is strongly word-by-word based on the content words (see Figure 4-2 and 4-3). However, consistent with the finding from the previous study, the earliness of constraint on the complement semantics in the brain was replicated using the LSTM network. The blended semantic effect that captures the combined representation of 50 possible candidates in 100 dimensional latent semantic (topic) space showed a significant model-fit in RHG around 300ms before the verb-onset. No other significant fit was observed in the other regions in contrast to the full-context (blended) semantic constraint based on the behavioural pretest data. This could possibly reflect the fact that the top 50 likely complement nouns suggested by the LSTM network is strongly based on the lexical semantics of a preceding word. However, other anterior temporal and inferior frontal regions (RaITG, RTP and RBA44) showed significant fits to a pure LSTM prediction model at a later stage of processing from 100ms before to 60ms after the verb-onset. The similarity pattern of this pure prediction model is based on the entire word-level prediction in a form of a probability distribution, not just based on combination of few most likely words. From these results, it is plausible that the early constraint on the complement is abstracted and is represented by RHG which becomes more specific as the subject noun is recognized in RaITG, RTP and RBA44.

Somewhat surprisingly, the LSTM prediction failed to capture the spatiotemporal dynamics of the anterior LIFG, the regions that represented the blended constraint based on the behavioural pretest. As discussed in Chapter 3, this region is involved in semantic ambiguity resolution, selection and unification, which contributes to more efficient processing of the upcoming word. In contrast, the blend model based on behavioural pretest data showed comparable effects in RTP with the pure LSTM prediction model. From these results, I suggest that the specific prediction strength (probability) for each individual candidate in the LSTM prediction does not reflect the neural prediction although the overall pattern across the prediction distribution (e.g. the clustering of candidates and their semantic/syntactic relations) tends to be similar; recall that the semantic blend is generated by weight-combining the topic representation of 50 most likely continuations which requires the prediction strength (probability) to be accurate. To test this hypothesis, I ran an additional analysis (not shown) using Euclidean distance instead of cosine distance to quantify the dissimilarity between a pair of prediction distributions which returned significant model-fits in none of the ROIs; recall that Euclidean distance is sensitive to the amplitude of each dimension whereas cosine distance captures the difference in orientation (i.e. the overall pattern in the distributions) unless the distributions are L2-normalized. This possibly explains why these effects around the verb-onset are only significant for the pure LSTM prediction model constructed by cosine distance.

Lastly, it is worth noting that RaSTG activation was consistently observed for constraining an upcoming verb and complement noun. In addition to constructing the early semantic constraint on the complement, this region was also involved in constraining the upcoming verb and complement noun captured by the pure LSTM prediction models. Further, this region was also consistently activated for lexical semantic processing captured by both the DM co-occurrence semantic model and the first internal layer of the LSTM network. In summary, this region is likely to be a central hub of predictive processing that simultaneously represents the lexical-semantic information of an input word and the lexical-semantic constraint on the upcoming input. Further, this region may bridge the interaction between the posterior (involved in lexical-semantic activation centred on pSTG) and anterior temporal areas (involved in computation of constraints centred on RTP). Future research could investigate the directed connectivity pattern between these regions during a spoken sentence comprehension to support this conjecture.

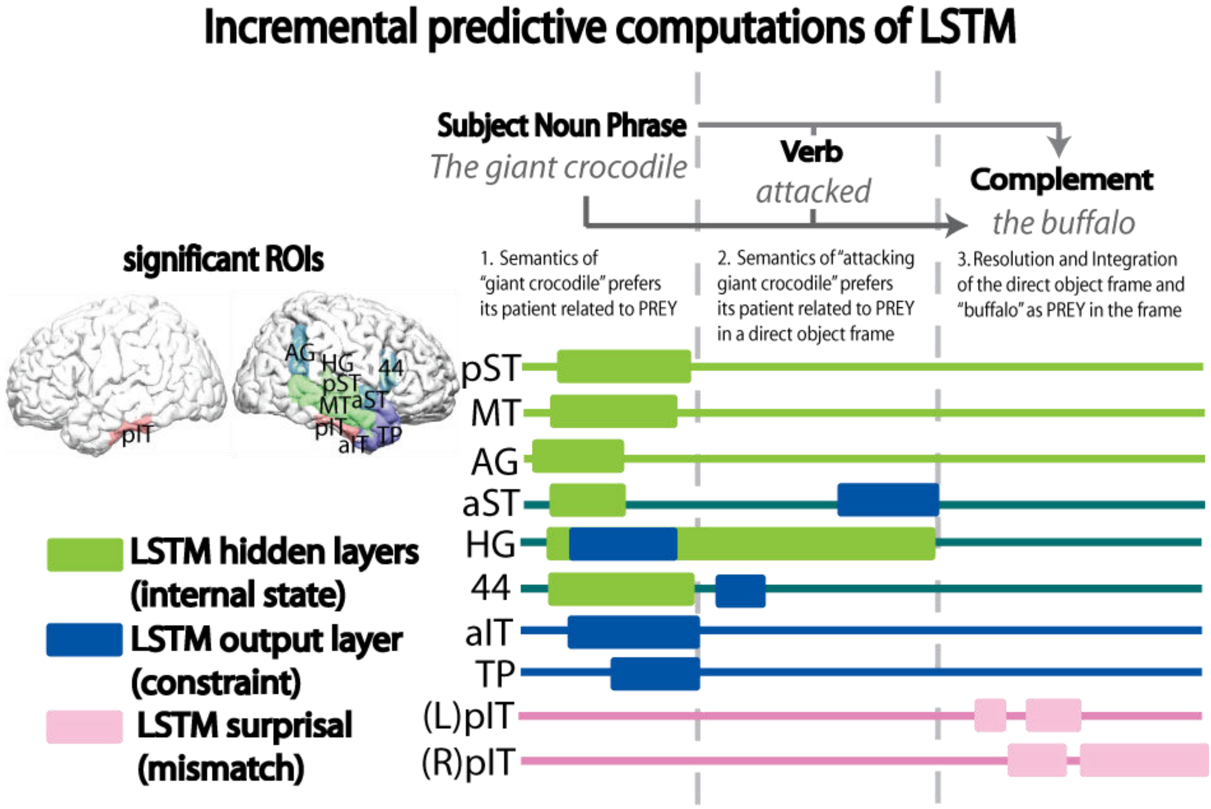

Incremental predictive computations of LSTM
Subject Noun Phrase
The giant crocodile
Verb
attacked
Complement
the buffalo
significant ROIs
1. Semantics of "giant crocodile" prefers its patient related to PREY
2. Semantics of "attacking giant crocodile" prefers its patient related to PREY in a direct object frame
3. Resolution and Integration of the direct object frame and "buffalo" as PREY in the frame
AG
HG
44
pST
MT
aST
pIT
aIT
TP
pST
MT
AG
aST
HG
44
aIT
TP
(L)pIT
(R)pIT
LSTM hidden layers (internal state)
LSTM output layer (constraint)
LSTM surprisal (mismatch)

Figure 4-14: Summary of results in the bilateral language network. RSA effects of LSTM computations in relation to prediction and integration during language comprehension. The effects of the hidden layers' and the output layer's computations are summarized in green and blue respectively. Further, the surprisal effects given the LSTM output are also presented in pink. The relative timing of each effect is shown by a bar(s) on the line that represents each region.

4.6.3. Integration

The aim of the additional analysis was to corroborate the facilitatory role of the constraint on processing of a subsequent input. The surprisal model of the complement noun from the LSTM prediction showed a significant correlation with LpITG as early as 100ms which peaked around 200ms after the noun onset. This region is commonly reported "semantic-processing" region (Binder et al., 2009) involved in semantic ambiguity resolution (Rodd et al., 2005). Just like resolving the ambiguity to interpret a word with a particular meaning, resolving the mismatch between the predicted and the actual target may involve this region as soon as the phonetic information becomes available. This mismatch resolution at the lexical level will lead to integration of the semantics of the target into the message-level representation in LpMTG as shown in the previous study in Chapter 3. Further, this analysis found that the first internal layer at the point of a complement noun (representing the semantics of the target and the context) characterizes the activity pattern in RITG from 170ms to 500ms after the noun onset. From these bilateral ITG regions, I suggest that the early lexical-level mismatch resolution triggers the activation of lexical semantic information which, in turn, will lead to the integrated representation at the message level.

4.6.4. Methodology: insights on the different modelling aspects

In conclusion, the current LSTM network lacks an important aspect of human speech comprehension: a non-adjacent word in the context could determine the constraint on the upcoming input. In fact, the LSTM architecture is attractive because it allows the network to capture long-distance dependencies in sentence processing. Inside the memory cell, this network combines the input embeddings with the memory content at the previous time-point, simply by calculating the weight-summation between these two vectors without any non-linear function (instead, the weights are computed from a non-linear transformation of a linear combination of three different components including the input embeddings, the previous memory content and the previous cell state). This architecture ensures that the error gradient is not reduced during the back-propagation through time (BPTT) training process, since the derivative of an identity function is still an identity; hence, the error gradient will not be reduced even if it passes through many recurrent projections over time. This is how this network solves the vanishing gradient problem and better explains the long-distance dependencies during sentence processing.

However, as demonstrated in 4.3, it is clear that the LSTM prediction of a complement noun tends to be more dependent on a preceding verb (adjacent content word) than a preceding subject noun (a distant content word). The correlations between the semantic constraint model and the second internal layer as well as the output layer models consistently increased at the point of a complement function word (see Figure 4-2 and 3-3), in contrast to the brain which constructs the semantic constraint as early as the subject noun. The main criticism can focus on the task that the network performed: this network is merely a predictive machine that tries to predict the upcoming word as accurately as possible. However, the ultimate goal of speech processing in humans is to understand the message that a speaker intends to convey which is learned through experience (e.g. talking with parents, playing with siblings etc.) without having an explicit task. Further, the actual gradient decent algorithm used in

Jozefowicz et al. (2016) is known as adaptive gradient optimizer (ADAGRAD) which adapts the learning rate parameter by normalizing the squared sum of the past gradients with respect to a weight in order to draw the network's attention to infrequent features (or rarely activated neurons). An issue with this algorithm is that the squared sum of past gradients is accumulated over time, making the algorithm rapidly diminish the error gradient (or teaching material) during the back-propagation of the gradient over a number of words in the context.

It is important to note that the lexical prediction of a complement function word did not model the syntactic constraint on the complement in the brain. It may potentially require the network to utilize syntactic knowledge more explicitly since the current network only implicitly activates syntax learned from word statistics in the massive corpora. This includes changing the task from predicting a lexical item to predicting a syntactic structure which will enable us to better capture the syntactic constraint explicitly in the form of a probability distribution. A recent study on the LSTM network by Linzen, Dupoux and Goldberg (2016) already demonstrated a miserable performance of the LSTM model by Jozefowicz et al. (2016) on multiple syntax-sensitive tasks (e.g. grammaticality judgment, number agreement and verb inflection) as a number of attractors (clauses in-between the main subject noun and the main verb) increased; its performance was not much different from a random guess when there were two or more attractors. As discussed above, this suggests that the LSTM network model in this study does not capture long-distance dependencies/hierarchical syntax because of its architecture.

Trends in language modelling: room for improvements

A potential solution is to introduce multi-task learning (MTL) paradigm in which more than one task is used to train a network (Liu, Su, Jia, Gao, Hao & Yang., 2015). In Liu et al's paper, a number of different LSTM architectures designed for MTL (uniform- vs. coupled vs. shared-layer architectures) are discussed in relation to their performance in text classification tasks. Secondly, a recent learning paradigm allows the network to map an input to an output sequence through the encoder-decoder framework (Cho, Van Merrienboer, Gulcehre et al., 2014). This paradigm enlightens how a network can process language at phrasal or sentential level; a model architecture and training paradigm for sentence-level representations were recently proposed by Kiros, Zhu, Salakhutdinov et al. (2015), known as skip-thought vectors. The encoder output from this framework directly represents a sentence embedding which is fed as an input to the decoder for output generation; see Luong, Sutskever, Vinyals and

Kaiser (2015) for supervised sequence learning with various settings. Lastly, another branch of language modelling networks introduces an efficient way of incorporating the hierarchical information explicitly represented as a parse tree (Zhu, Sobihani & Guo., 2015). This framework incorporates a binary parse tree into the formulation by adding hidden $(h_{t-1}^{L}, h_{t-1}^{R})$ and cell vectors $(c_{t-1}^{L}, c_{t-1}^{R})$ gated with separated forget gates, assuming that each tree node can only have two children underneath with multiple descendants. Whether to pass or block information from a node is determined by sigmoid weights trained through a corpus. From these frameworks, a network can evolve from a strictly word-by-word processing machine to a machine that utilizes and processes various structures from a word to a sentence, which could make the network a better descriptive model of human speech comprehension.

Chapter 5: General discussion

The central issue investigated in this thesis concerns the temporal neurodynamics of the incremental computations involved in speech comprehension across the brain. By constructing a number of different models of linguistic constraints on the upcoming language input and testing them against the spatiotemporal dynamics of neural activity, the predictive nature of human speech comprehension was corroborated where the full-context constraint on the semantics of the complement phrase was initially activated around the time when lexical semantics of a subject noun was activated. However, consistent with lexical functional grammar (Bresnan, 1981), a verb's lexico-syntactic SCF constraint on the complement structure (regardless of whether it's verb-based or full-context) showed effects specifically after the verb onset, around when the verb was recognised. The early activation of the semantic constraint generated by the SNP was also replicated by an LSTM network model learned from word-level statistics, although the syntactic constraint effects were not observed. In this chapter, I will discuss a more detailed neurobiological account of incremental speech comprehension by bringing the results from the previous chapters together and evaluating them against previous research and psycholinguistic accounts. The temporal progression of linguistic predictive information and the contribution of these studies to understanding the neurobiological basis of incremental speech comprehension will be highlighted and the five questions stated in Chapter 1 will be addressed.

5.1. Advantages of computational modelling in explaining neurobiological data

The majority of natural linguistic variables are probabilistic and reflect our experience of encountering language in the world. Building plausible models to explore various linguistic phenomena is, thus, a necessary step to understanding language processing in humans. Owing to recent technological developments, corpus linguistics has attracted attention from researchers from various fields of applied linguistics. It is based on a massive set of language samples which allows researchers to analyze various aspects of language. In this way, this approach provides a number of highly reliable and objective models each of which can capture a particular aspect of linguistic computations. In conjunction, connectionist theories have also gained much attention because a neurobiologically inspired machine (a.k.a. neural network) can be trained through the large corpus, learning from non-linear statistical patterns across a massive number of language samples to generate an accurate response for a task (e.g.

predicting an upcoming word). The important advantage of using a neural network model to study human language comprehension is that it shows how each incrementally unfolding word changes the current state of the system from which the subsequent prediction is generated. Therefore, this approach is particularly attractive because the “incrementality” of linguistic computations during speech comprehension can be explored. Consistent with the conclusion in Chapter 4, a previous study by Wehbe et al. (2014) have used an RNN model and showed that the RNN model’s 1) word embedding, 2) internal layer and 3) output probability can be used to predict the MEG data. However, this field is relatively young and developing biologically plausible neural network models is still an ongoing research topic (Bengio, Lee, Bornschein, Mesnard & Lin, 2015). There are many different ways to improve a neural network as a descriptive model of human speech comprehension; note that some limitations of the state-of-art neural network model used in this thesis (Jozefowicz et al., 2016) for modelling human speech comprehension are discussed in Chapter 4.

Investigating which linguistic variables explain linguistic phenomena involves testing to see how much each variable co-varies with the response measure. In this way, one could statistically test the relations between various linguistic properties of each linguistic unit and the human response measure. However, some linguistic variables are defined in a multidimensional space including predictions which are modelled in the form of a probability distribution among different candidates. Similarly, unlike a behavioural response measure such as reaction time, the neurobiological data naturally varies over space and time, representing dynamic patterns of response. This is why a multivariate data analysis approach is motivated to characterize the cognitive processes in human brain and, RSA in particular, provides a way to relate the pattern of information encoded in neural activity and in computational models with varying number of dimensions. In this way, the neuro-cognitive processes characterized by a set of changing representational information over space and time can be investigated through modelling the information in the representational space defined by each neuron over space and time. Using this approach, the analysis avoids losing variability in the original data space unlike the traditional approaches of summarizing the neural activity into the univariate amplitude (either by averaging or by finding a first eigenvariate for each ROI).

5.2. Predictive processing in incremental speech comprehension

Processing a word in a sentence involves incremental computations relating each word to the preceding context. In predictive accounts of human language comprehension, the human brain utilizes contextual constraints to facilitate the processing of a word in a sentence, as consistently shown in this thesis and elsewhere (see Kuperberg & Jaeger, 2016). According to Kuperberg (2016), listeners undergo a series of incremental computations of predicting an upcoming input and updating the context once the input is heard in order to effectively infer the event from a set of hierarchically organized representations (see Figure 2-2). These representations allow listeners to evaluate the language input and its statistical properties based on the beliefs about the message that a speaker intends to convey. Using the models of multi-level constraints and their error with respect to the actual input, this thesis investigated the neurobiological basis of the incremental computations which involve a series of predictions and updates throughout a sentence. By characterizing the response patterns of neural activity using these models, the overall findings from this thesis consistently reported the significant effects of constraints followed by the effects of error at multiple linguistic levels, reflecting the cognitive process of lexical, syntactic and semantic predictions and updates (integration) during incremental speech comprehension.

What are the linguistic bases of predictive computations?

Based on the predictive nature of incremental speech processing that has been firmly established in the literature (Altmann & Mirkovic, 2009; Federmeier & Kutas, 2011; Delong et al., 2014; Kuperberg & Jaeger, 2016), this thesis explored the linguistic bases of predictive computations at syntactic and semantic levels. Previous studies have consistently found that the subcategorization frame (SCF) preference of a verb plays an important role in constraining the syntactic interpretation of its complement (Trueswell et al., 1993; Jennings et al., 1997; Gibson & Pearlmutter, 1998). Supporting this argument, I showed in Chapter 3 that both syntactic (SCF) constraints based on a verb and on the full preceding context were activated in left lateralized fronto-temporal regions soon after the onset of a verb (around 170ms). In contrast, semantic constraint was exclusively based on the full preceding context and none of the models of lexico-semantic constraint showed significant effects in any of the ROIs. This absence of verb-based constraint effects on the complement semantics does not support the lexicalist claim.

At a first glance, the absence of lexical semantic effects is somewhat surprising given the evidence from previous studies that a verb directly constrains the semantic/pragmatic properties of its argument (Marslen-Wilson et al., 1988; Hare et al., 2003; Bicknell et al., 2010). However, Nieuwland and Van Berkum (2006) showed that local semantic/pragmatic constraint is strongly influenced by the context in which it is presented. In line with this finding, Kamide, Altmann and Haywood (2003) demonstrated that a pre-verbal argument (agent) constrains the subsequent theme in combination with a verb. They also showed that the pre-verbal argument constrains the forthcoming arguments in Japanese (which is an example of head-final language), demonstrating that a verb is not the only driving factor of predictive processing. These studies offer an alternative interpretation as follows; the absence of verb-based semantic constraint effects in this thesis is likely because the rich subject NP (e.g. “The experienced walker”) provides stronger constraints in general on the complement semantics (e.g. “the path”) such that a verb (e.g. “chose”) only plays a confirmatory role during the predictive processing. Consistent with this interpretation, the ongoing study, which minimized the contextual influence of the subject NP, observed the late effects of verb-based semantic constraints around the uniqueness point of a verb while replicating the full-context semantic constraint effects in the bilateral fronto-temporal regions before the verb-onset.

Regardless, the finding that semantic constraint is strongly based on the entire preceding context before the verb-onset suggests that the incremental predictive computations in humans are driven by the combined properties of lexical constraints such that each lexically-driven constraint is modified by the preceding context if it exists. This is consistent with the lexicalist accounts claiming that the content of each upcoming word is constrained, evaluated and integrated into the context (Marslen-Wilson, 1975; Marslen-Wilson & Tyler, 1980; Marslen-Wilson et al., 1988; Sag & Wasow, 2011). The different linguistic bases on syntactic and semantic constraints directly address another question below.

Are syntactic constraints activated prior to the activation of semantic constraints in order to enable early phrase structure building before constraining the lexical-semantics?

In this thesis I found that syntactic constraint differs from semantic constraint in the context on which it is based as discussed above. As a result, the semantic constraint effects appeared soon after the onset of an initial subject noun whereas the syntactic constraint effects emerged only about 170ms after the onset of a verb. This finding suggests that the explicit phrase structure building is not a necessary requirement for constraining the complement semantics. Instead, it can be constrained as soon as thematic information of a subject is revealed. This is not consistent with the syntax-first theory (Frazier, 1978; 1987; Friederici, 2002) which emphasizes the initial stage of phrase structure building independent of lexical semantics, which is only activated at the later thematic assignment stage once the syntactic structure is built. Throughout Chapter 3 and 4 in this thesis, the predictive nature of incremental speech comprehension is demonstrated in which listeners constrain the upcoming input based on contextual properties and semantic constraints are consistently activated before the syntactic constraints.

In contrast to the ERP-based evidence supporting the syntax-first theory, in this thesis I analyzed the source-localized EMEG data, recorded while participants were listening to natural speech, using the state-of-art computational models of predictive processing. Taking a multivariate pattern analysis approach allowed the brain's response patterns to be characterized using the rich multidimensional information encoded in predictive computations with millisecond resolution. Hence, the results in this thesis are improved in three different aspects compared to the classical ERP studies; 1) given that the stimuli are all natural sentences without violations, they can be more reliably generalized to natural speech comprehension, 2) they highlight the regions and networks involved in different linguistic computations from which the underlying neural mechanism can be elucidated 3) they present the temporally specific effects with high temporal resolution and do not suffer from the consistency issue in interpreting the results due to summarizing the effects over a large time-window.

In summary, these results partly support the lexicalist account claiming that both syntactic and semantic information is localized within lexical entries from which constraints are constructed (Sag & Wasow, 2011) and fully consistent with the parallel-interaction theory suggesting that multiple linguistic aspects of the context interact and provide maximally incremental interpretation of an upcoming speech (Marslen-Wilson, 1975; Tyler & Marslen-Wilson, 1977; Marslen-Wilson & Tyler, 1980). In particular, they emphasize that semantic constraints are more flexibly constructed such that a verb-based constraint can be overshadowed by the thematic constraint from a subject because a verb cannot account for the thematic association between a preceding subject and an upcoming object.

Utilizing constraints for integration

In order to ensure that the activated constraints are applied to facilitate the processing of a target word, in my analyses the amount of error in the constraints was quantified and tested against neural activity after the onset of the target word. Quantifying such error is another important computation to obtain a converged event representation through minimizing the unexplained proportion of variance in predictions by the bottom-up input (Kuperberg, 2016). Here, the amount of error (e.g. surprisal) directly captures the amount of cognitive effort to integrate a word into the context (Hale, 2001; Levy, 2008) which has been commonly used as an index of linguistic integration and captured the variability in human responses to different target words with varying degree of error with respect to the context (Roark et al., 2009; Frank & Bod, 2011; Fossum & Levy, 2012; Smith & Levy, 2013). In addition to these studies, this thesis showed that neural activation of constraints before the onset of a word is generally followed by representations of the constraint error after the word's onset. For example, in Chapter 3, the semantic constraint error was significantly represented in LpMTG between 280 and 600ms after the target word onset, which was around 100ms after the constraint effect declined in L-BA45. Similarly, the effect of syntactic constraint which declined around 530ms after the verb-onset in L-BA44 was followed by the error response around 170ms after the target word onset in L-BA45 (which was, on average, 100 – 150ms after the constraint effect declined). These results clearly demonstrated the facilitatory role of constraints on processing a word.

Are predictive processes of human speech comprehension based on explicit statements of syntactic rules?

Within the predictive framework of speech comprehension, the rule-based account of human speech comprehension has attracted considerable attention. It claims that predictions are based on nested syntactic structure rather than a sequence of words. This account has recently been brought into focus by Ding et al. (2016) showing that neural activity is entrained to the frequency of the stimulus presentation at syllabic (1Hz), phrasal (2Hz) and sentential (4Hz) levels. This result was interpreted as evidence for cortical tracking of hierarchical structures, claiming that the statistical relationships between words alone cannot sufficiently explain human speech comprehension (see Ding et al., 2017).

As a response to this study, Frank and Yang (2018) replicated this result only using word-level statistics and claimed that understanding a sentence with simple syntactic structure can be achieved from the statistical information associated with each word without applying syntactic rules. Consistent with this claim, the models of syntactic and semantic constraints based on co-occurrence statistics showed significant correlations with neural activity at different points in a sentence. Especially, the significant syntactic constraint effects around 170ms after the verb-onset imply that the syntactic understanding of a sentence can be driven by activating the co-occurring structures with the verb without rule-based analysis of phrasal construction, at least for a simple grammatical sentence. Additionally, this syntactic constraint model was significantly correlated with the network's prediction on the verb's complement, which does not have any explicit knowledge of syntax, demonstrating the lexical nature of predictive processing in human speech comprehension. Although these results are consistent with Frank & Yang (2018), future research should investigate the degree to which these results can be generalized to processing more syntactically complex sentences which include long-distance dependencies.

How incremental is the predictive processing in human speech comprehension?

Whilst incrementality is the key property of speech comprehension, the degree of incrementality in predictive processing is less clear. This question was directly addressed in this thesis by investigating the similarities and differences between the computations involved in the brain and the network model trained to predict every incrementally unfolding word in a sentence. The results showed that the network's internal processing states and output prediction significantly capture the response patterns of the brain only for the content words including a subject noun, a verb and a complement noun, but not a function word that indicates the syntactic structure of the complement. Interestingly, none of the network layers at the point of a verb characterized the neural response patterns, suggesting that the network's computations that predict the upcoming function word were not consistent with the neurobiological computations that predict the syntactic structure of the complement in the left fronto-temporal network. Taken together, it can be suggested from these results that syntactic constraint in humans is not as specific as in the network and only predicts the words that are semantically meaningful to construct the event representation.

5.3. Neurobiological account of syntax and semantics in predictive computations for incremental speech comprehension

Exploring the similarities and differences between syntactic and semantic processes in human language comprehension has long been a topic of interest in the field of neuroscience. The majority of neurobiological accounts agree that syntax recruits more left-lateralized fronto-temporal network whereas semantics elicits greater activity in the bilateral fronto-temporo-parietal network (Tyler & Marslen-Wilson, 2008; Tyler et al, 2010; Price, 2010, 2012; Friederici, 2011; Hagoort, 2013). Consistent with these accounts, this thesis replicated this functional distinction between syntax and semantics that specifically represent the predictive properties on an upcoming complement phrase.

Distinction between right ATL and left ATL in time

One of the consistent findings throughout this thesis is that the right anterior temporal regions are activated for the early construction of the constraint on the complement noun at the point of a subject noun. For example, the full-context semantic constraint in Chapter 3 showed a significant effect in RaSTG soon after the onset of a subject noun (on average), followed by an effect in RTP peaking around the verb onset. Similarly, the LSTM network's prediction on the complement noun was reflected in RaITG soon after the onset of a subject noun and peaked around 100ms before the verb-onset together with RTP (see Chapter 4). Previous neurobiological accounts suggest that the bilateral ATL is typically involved in combinatorial processing at both syntactic and semantic levels during natural language comprehension (Hickok & Poeppel, 2007; Rogalsky & Hickok, 2008). Consistent with these findings, Bornkessel-Schlesewsky and Schlesewsky (2013) suggested a role of this region for time independent processing of building and unifying/combining the conceptual schemata to track and develop a sentence-level representation. In addition to these claims, the findings in this thesis suggest that developing a sentence-level representation in this region naturally leads to constraining the subsequent themes based on the theme of a subject NP (agent). This computation is central to the early stage of predictive processing in a sentence to facilitate the understanding of incrementally unfolding words in a semantically coherent manner at the sentence-level.

Unlike the claim that such combinatorial processing involves bilateral ATL, the early representation of the semantic constraint around the subject noun only recruited right ATL.

However, this right ATL effect was followed by a marginally significant effect in left temporal pole (LTP) around the time in which the semantic constraint was activated in L-BA47 after the verb is recognized (around 330ms after the verb-onset). Therefore, once a verb confirms the early semantic constraint constructed by a subject NP, the representation of this constraint weakly appears in the left homolog region. Given the significant effect in L-BA47 around the same time, it is likely that the early constraint based on the subject NP at this point in time is semantically unified with the verb-based constraint (Hagoort, 2013) for selecting the likely candidates more specifically. Consistent with this interpretation, Jung-Beeman (2005) claimed that temporal regions in the right hemisphere (RH) represent more coarse-grained semantics with larger and more diffused semantic fields. In particular, he suggested the role of RTP in computing the degree of semantic overlap among the coarse-grained semantic fields to support message level interpretation. Taken together, these results are consistent with the previous neurobiological accounts that bilateral ATL is involved in combinatorial processing to develop a sentence-level representation, but additionally highlight that 1) subsequent themes are naturally constrained from the sentence-level representation and 2) right ATL is engaged in constructing the semantic constraint at the early stage in which the candidate themes are semantically general and coarse-grained whereas left ATL (possibly through interaction with L-BA47) represents an unified constraint to make it more specific and fine-grained. This relationship between bilateral ATL in time is particularly informative as it has never been explained by the previous neurobiological models due to the lack of EMEG evidence having high temporal resolution.

Multiple functional roles of left MTG/ITG

In Chapter 3, the pattern of activity in left MTG was significantly correlated with the lexico-syntactic constraint of a verb from around 170ms after the verb-onset. The importance of this region in syntactic processing is consistently found by previous studies (Tyler, Stamatakis, Post, Randall & Marslen-Wilson., 2005; Rodd et al., 2010). Especially, a previous study which manipulated the syntactic ambiguity of a subject NP (e.g. “juggling knives”) showed that a direct object preference of a verb in the phrase (e.g. “juggle”) is represented in the posterior portion of LMTG around the offset of the verb (Tyler et al., 2013). Consistent with these findings, Hagoort (2013) suggested the role of this region in accessing the lexico-syntactic information from memory which is unified in LIFG for selective pre-activation (Snijders et al., 2008, 2010).

On the other hand, this region was also observed for representing the error in the semantic constraint from 280ms after the onset of a complement noun, consistent with the claim that this region is involved in lexical-semantic access (Hickok & Poeppel, 2007). Given that this region and timing is where N400 is typically localized (Simos et al., 1993), the error likely reflects the ease with which lexical information of a target is accessed (Lau et al., 2008). An alternative interpretation suggests that the amount of error (e.g. surprisal) directly captures the amount of cognitive effort to integrate a word into the context (Hale, 2001; Levy, 2008) and has been commonly used as an index of linguistic integration in psycholinguistic research (Roark et al., 2009; Frank & Bod, 2011; Fossum & Levy, 2012; Smith & Levy, 2013). The early lexical access in sentence processing (Hauk & Pulvermuller, 2004) supported this interpretation.

Consistent with this interpretation, it was further shown that the error in the LSTM prediction was represented in LpITG around 190ms after the complement noun onset; the region involved in activating the lexico-semantic properties (Hickok & Poeppel, 2007; Bingjiang et al., in prep) and resolving semantic ambiguities (Rodd et al., 2005). The emergence of this effect in earlier time-window which is more transient than the semantic constraint error possibly reflects a quick integration of the lexical form of the noun before unifying its semantics into a sentence-level representation. Supporting this argument, this effect was followed by the RITG activity reflecting the LSTM network's internal state (HL06) at the point of a complement noun. Since the internal state represents the weighted combination between the context representation (captured by the previous memory contents) and the lexical embeddings of a current input, its representation essentially reflects the integrated properties from which the network's subsequent prediction is constructed.

Taken together, these results imply that left MTG/ITG regions play multiple functional roles during predictive processing of a spoken sentence including 1) activating lexico-syntactic constraint, 2) activating lexico-semantic constraint (shown in Bingjiang et al. in prep) and 3) utilizing these constraints to facilitate the semantic processing of a target word. However, in order to corroborate the semantic integration account, future research must explore the neurally plausible function of semantic composition to directly test changes in the semantic representation before and after integration (e.g. see Hartung, Kaupmann, Jebbara & Cimiano, 2017; Garten, Sagae, Ustun & Dehghani, 2015). Also, further functional connectivity studies should clarify the way that these regions interact with the other regions in the extensive language network for various predictive computations at syntactic and semantic levels.

Constraints utilization and LIFG

Processing a word in a sentence is clearly different from processing a word in isolation. In predictive accounts of human language comprehension, the human brain utilizes contextual constraints to facilitate the processing of a word in a sentence, as consistently shown in this thesis and elsewhere. The left inferior frontal gyrus (LIFG) has consistently been reported as a region that interactively process a lexical item with the auditory temporal regions by applying prior expectations for selection and integration (Tyler & Marslen-Wilson, 2008; Zhuang et al., 2012, Tyler et al., 2013; Kocagoncu et al., 2017; Cope, Sohoglu, Seddley et al., 2017). Moreover, the involvement of the LIFG also occurs when there is no experimental task. For example, Klimovich-Gray et al. (2019) have shown effects of LIFG in a study in which participants listened to two-word phrase stimuli with varying strength of semantic constraints of a first word (modifier) on a second word (noun) in a task-free environment. They found significant competition (entropy) effects of the constraint in L-BA45 starting around 70ms before the modifier offset and lasting until 165ms after the noun onset.

The same pattern of results was observed for both syntactic and semantic constraints in this thesis. For example, a transient effect of the semantic constraints was observed in L-BA45 around 150ms after the complement noun onset (see Figure 3-5(a)). Similarly, lexical syntactic (SCF) constraints were represented in L-BA44 which declined around the offset of the complement function word which directly indicates the complement structure (see Figure 3-6 panel B). Around 100-150ms after these effects disappeared, the representations of constraints error for syntax in L-BA45 and for semantics in LpMTG emerged as discussed above. In summary, this thesis using source-localized EMEG data supports the role of LIFG in predictive processing for utilizing constraints to facilitate the bottom-up processing of a target word during incremental speech comprehension which shows a functional distinction between applying different levels of constraints (i.e. semantic-anterior and syntactic-posterior), consistent with previous neurobiological models of speech comprehension (Hagoort, 2005, 2013; Bornkessel-Schlesewsky & Schlesewsky, 2013).

5.4. Conclusion

In conclusion, consistent with the lexicalist account of speech comprehension, the predictive computations in the brain involves activating multi-level constraints and utilizing them to facilitate the processing of a target word. Nevertheless, the time at which these constraints are activated varied. For example, syntactic constraint is strongly driven by the lexical property of a verb (i.e. SCF) which appears strictly after the verb onset whereas semantic constraint is based more strongly on the preceding subject NP and emerges soon after the onset of a subject noun. In particular, a preceding theme strongly constrains the subsequent themes so that individual words can be interpreted in a semantically coherent fashion with respect to a message-level representation. These predictive processes are incremental; each (content/meaningful) word in a spoken sentence changes the state of the brain from which constraints on the subsequent input are computed. In this way, the brain actively predicts and integrates a number of themes throughout a sentence and reaches at the converged representation of a message.

References

Ahlfors, S. P., Han, J., Belliveau, J. W., & Hämäläinen, M. S. (2010). Sensitivity of MEG and EEG to source orientation. *Brain Topography*, *23*(3), 227-232.

Altmann, G. (1988). Ambiguity, parsing strategies, and computational models. *Language and Cognitive Processes*, *3*(2), 73-97.

Altmann, G. T., & Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science*, *33*(4), 583-609.

Altmann, G., & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition*, *30*(3), 191-238.

Altmann, G., & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition*, *30*(3), 191-238.

Altmann, G., & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition*, *30*(3), 191-238.

Anwander, A., Tittgemeyer, M., von Cramon, D. Y., Friederici, A. D., & Knösche, T. R. (2006). Connectivity-based parcellation of Broca's area. *Cerebral Cortex*, *17*(4), 816-825.

Baayen, R.H., Piepenbrock, R., and Guilikers, L. (1996). CELEX2 Database (CD-ROM). Linguistic Data Consortium, http://www.ldc. upenn.edu/.

Barnett, L., & Seth, A. K. (2014). The MVGC multivariate Granger causality toolbox: a new approach to Granger-causal inference. *Journal of Neuroscience Methods*, *223*, 50-68.

Baroni, M., & Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, *36*(4), 673-721.

Bates, E., Wilson, S. M., Saygin, A. P., Dick, F., Sereno, M. I., Knight, R. T., & Dronkers, N. F. (2003). Voxel-based lesion–symptom mapping. *Nature Neuroscience*, *6*(5), 448.

Beeman, M., Friedman, R. B., Grafman, J., Perez, E., Diamond, S., & Lindsay, M. B. (1994). Summation priming and coarse semantic coding in the right hemisphere. *Journal of Cognitive Neuroscience*, *6*(1), 26-45.

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, *7*(6), 1129-1159.

Bever, T. G. (1970). The cognitive basis for linguistic structures. *Cognition and the Development of Language*, *279*(362), 1-61.

Bicknell, K., Elman, J. L., Hare, M., McRae, K., & Kutas, M. (2010). Effects of event knowledge in processing verbal arguments. *Journal of Memory and Language*, *63*(4), 489-505.

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, *19*(12), 2767-2796.

Binder, J. R., Frost, J. A., Hammeke, T. A., Cox, R. W., Rao, S. M., & Prieto, T. (1997). Human brain language areas identified by functional magnetic resonance imaging. *Journal of Neuroscience*, *17*(1), 353-362.

Binder, J. R., Frost, J. A., Hammeke, T. A., Bellgowan, P. S., Springer, J. A., Kaufman, J. N., & Possing, E. T. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex*, *10*(5), 512-528.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993-1022.

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2013). Reconciling time, space and function: a new dorsal–ventral stream model of sentence comprehension. *Brain and Language*, *125*(1), 60-76.

Bornkessel, I., Zysset, S., Friederici, A. D., Von Cramon, D. Y., & Schlesewsky, M. (2005). Who did what to whom? The neural basis of argument hierarchies during language comprehension. *Neuroimage*, *26*(1), 221-233.

Brennan, J., Nir, Y., Hasson, U., Malach, R., Heeger, D. J., & Pylkkänen, L. (2012). Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and Language*, *120*(2), 163-173.

Bresnan, J. (1981). An approach to Universal Grammar and the mental representation of language. *Cognition*.

Bresnan, J. (2001). *Lexical-Functional Syntax*. Blackwell

Briscoe, T., & Carroll, J. A. (2002). Robust Accurate Statistical Annotation of General Text. In *LREC*.

Brown, C., & Hagoort, P. (1993). The processing nature of the N400: Evidence from masked priming. *Journal of Cognitive Neuroscience*, *5*(1), 34-44.

Campbell, K. L., & Tyler, L. K. (2018). Language-related domain-specific and domain-general systems in the human brain. *Current Opinion in Behavioral Sciences*, *21*, 132-137.

Caplan, D. (1999). Activating brain systems for syntax and semantics. *Neuron*, *24*(2), 292-293.

Caplan, D., Alpert, N., & Waters, G. (1998). Effects of syntactic structure and propositional number on patterns of regional cerebral blood flow. *Journal of Cognitive Neuroscience*, *10*(4), 541-552.

Catani, M., & Jones, D. K. (2005). Perisylvian language networks of the human brain. *Annals of Neurology*, *57*(1), 8-16.

Clarke, A., Taylor, K. I., Devereux, B., Randall, B., & Tyler, L. K. (2012). From perception to conception: how meaningful objects are processed over time. *Cerebral Cortex*, *23*(1), 187-197.

Colclough, G. L., Brookes, M. J., Smith, S. M., & Woolrich, M. W. (2015). A symmetric multivariate leakage correction for MEG connectomes. *Neuroimage*, *117*, 439-448.

Cope, T. E., Sohoglu, E., Sedley, W., Patterson, K., Jones, P. S., Wiggins, J., ... & Davis, M. H. (2017). Evidence for causal top-down frontal contributions to predictive processes in speech perception. *Nature Communications*, *8*(1), 2154

Coulson, S, King, J. W, and Kutas, M. (1998). "Expect the unexpected: Event-related brain response to morphosyntactic violations." *Language and Cognitive Processes*13.1: 21-58.

Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., & Robinson, T. (2013). One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Chomsky, N. (1964). *Aspects of the Theory of Syntax*. MASSACHUSETTS INST OF TECH CAMBRIDGE RESEARCH LAB OF ELECTRONICS.

Chomsky, N. (1981). Lectures on Government and Binding. Dordrecht: Foris Publications.

Chomsky, N. (1982). *Some Concepts and Consequences of the Theory of Government and Binding*. Linguistic Inquiry Monograph 6. MIT Press

Chomsky, N. (1993). *A Minimalist Program for Linguistic Theory*. MIT occasional papers in linguistics no. 1. Cambridge, MA: Distributed by MIT Working Papers in Linguistics

Chomsky, N. (1995). The Minimalist Program. Cambridge, Mass.: The MIT Press

Crain, S. (1980). Pragmatic constraints on sentence comprehension. Unpublished Ph.D. dissertation, University of California at Irvine.

Crain, S. & Steedman, M. J. (1985). On not being led up the garden path: The use of context by the psychological parser. In D. Dowty, L. Karttunen, & A. Zwicky (Eds), Natural language parsing: Psychological, computational, and theoretical perspectives. Cambridge: Cambridge University Press.

Crinion, J. T., Lambon-Ralph, M. A., Warburton, E. A., Howard, D., & Wise, R. J. (2003). Temporal lobe regions engaged during normal speech comprehension. *Brain*, *126*(5), 1193-1201.

Dale, A. M., & Sereno, M. I. (1993). Improved localizadon of cortical activity by combining EEG and MEG with MRI cortical surface reconstruction: a linear approach. *Journal of Cognitive Neuroscience*, *5*(2), 162-176.

DeLong, K. A., Troyer, M., & Kutas, M. (2014). Pre-processing in sentence comprehension: Sensitivity to likely upcoming meaning and structure. *Language and Linguistics Compass*, *8*(12), 631-645.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117.

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9-21.

Demonet, J. F., Chollet, F., Ramsay, S., Cardebat, D., Nespoulous, J. L., Wise, R., ... & Frackowiak, R. (1992). The anatomy of phonological and semantic processing in normal subjects. *Brain*, *115*(6), 1753-1768.

Démonet, J. F., Price, C., Wise, R., & Frackowiak, R. S. J. (1994). Differential activation of right and left posterior sylvian regions by semantic and phonological tasks: a positron-emission tomography study in normal human subjects. *Neuroscience Letters*, *182*(1), 25-28

Devereux, B. J., Tyler, L. K., Geertzen, J., & Randall, B. (2014). The Centre for Speech, Language and the Brain (CSLB) concept property norms. *Behavior Research Methods*, *46*(4), 1119-1127.

Ding, M., Chen, Y., & Bressler, S. L. (2006). Granger causality: basic theory and application to neuroscience. *Handbook of time series analysis: recent theoretical developments and applications*, 437-460.

Ding, N., Melloni, L., Tian, X., & Poeppel, D. (2017). Rule-based and word-level statistics-based processing of language: insights from neuroscience. *Language, Cognition and Neuroscience*, *32*(5), 570-575.

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, *19*(1), 158.

Dronkers, N. F., Wilkins, D. P., Van Valin Jr, R. D., Redfern, B. B., & Jaeger, J. J. (2004). Lesion analysis of the brain areas involved in language comprehension. *Cognition*, *92*(1-2), 145-177.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179-211

Elman, J. L. (2011). Lexical knowledge without a lexicon?. *The Mental Lexicon*, *6*(1), 1-33.

Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Research*, *1146*, 75-84.

Ferreira, F., & Clifton Jr, C. (1986). The independence of syntactic processing. *Journal of Memory and Language*, *25*(3), 348-368.

Fodor, J. A. (1983): The modularity of mind. Cambridge, MA: MIT Press

Formisano, E., Kim, D. S., Di Salle, F., van de Moortele, P. F., Ugurbil, K., & Goebel, R. (2003). Mirror-symmetric tonotopic maps in human primary auditory cortex. *Neuron*, *40*(4), 859-869.

Fossum, V., & Levy, R. (2012, June). Sequential vs. hierarchical syntactic models of human incremental sentence processing. In *Proceedings of the 3rd workshop on cognitive modeling and computational linguistics* (pp. 61-69). Association for Computational Linguistics.

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Van Essen, D. C., & Raichle, M. E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences*, *102*(27), 9673-9678.

Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, *22*(6), 829-834.

Frank, S. L., & Christiansen, M. H. (2018). Hierarchical and sequential processing of language: A response to: Ding, Melloni, Tian, and Poeppel (2017). Rule-based and word-level statistics-based processing of language: insights from neuroscience. Language, Cognition and Neuroscience. *Language, Cognition and Neuroscience*, 1-6.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2013). Word surprisal predicts N400 amplitude during reading. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Vol. 2, pp. 878-883).

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1-11.

Frank, S. L., & Yang, J. (2018). Lexical representation explains cortical entrainment during speech comprehension. *PloS one*, *13*(5), e0197304.

Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, *6*(4), 291-325.

Frazier, L. (1987). Sentence processing: A tutorial review.

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences*, *6*(2), 78-84.

Friederici, A. D. (2009). Pathways to language: fiber tracts in the human brain. *Trends in Cognitive Sciences*, *13*(4), 175-181.

Friederici, A. D. (2011). The brain basis of language processing: from structure to function. *Physiological Reviews*, *91*(4), 1357-1392.

Friederici, A. D. (2012). The cortical language circuit: from auditory perception to sentence comprehension. *Trends in Cognitive Sciences*, *16*(5), 262-268.

Friederici, A. D., Bahlmann, J., Heim, S., Schubotz, R. I., & Anwander, A. (2006). The brain differentiates human and non-human grammars: functional localization and structural connectivity. *Proceedings of the National Academy of Sciences*, *103*(7), 2458-2463.

Friederici, A. D., Pfeifer, E., & Hahne, A. (1993). Event-related brain potentials during natural speech processing: Effects of semantic, morphological and syntactic violations. *Cognitive Brain Research*, *1*(3), 183-192.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815-836.

Friston, K. (2008). Hierarchical models in the brain. *PLoS Computational Biology*, *4*(11). e1000211.

Friston, K., Harrison, L., Daunizeau, J., Kiebel, S., Phillips, C., Trujillo-Barreto, N. & Mattout, J. (2008). Multiple sparse priors for the M/EEG inverse problem. *NeuroImage*, *39*(3), 1104-1120.

Friston, K. J., Worsley, K. J., Frackowiak, R. S., Mazziotta, J. C., & Evans, A. C. (1994). Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*, *1*(3), 210-220.

Gagnepain, P., Henson, R. N., & Davis, M. H. (2012). Temporal predictive codes for spoken words in auditory cortex. *Current Biology*, *22*(7), 615-621.

Gallistel, C. R., Krishan, M., Liu, Y., Miller, R., & Latham, P. E. (2014). The perception of probability. *Psychological Review*, *121*(1), 96.

Garten, J., Sagae, K., Ustun, V., & Dehghani, M. (2015). Combining distributed vector representations for words. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing* (pp. 95-101).

Geschwind, N. (1965). Disconnexion syndromes in animals and man. *Brain*, *88*(3), 585-585.

Gers, F. A., & Schmidhuber, J. (2000). Recurrent nets that time and count. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium* (Vol. 3, pp. 189-194). IEEE.

Gibson, E., & Pearlmutter, N. J. (1998). Constraints on sentence comprehension. *Trends in Cognitive Sciences*, *2*(7), 262-268.

Gorno-Tempini, M. L., Dronkers, N. F., Rankin, K. P., Ogar, J. M., Phengrasamy, L., Rosen, H. J., ... & Miller, B. L. (2004). Cognition and anatomy in three variants of primary progressive aphasia. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, *55*(3), 335-346

Gorno-Tempini, M. L., Rankin, K. P., Woolley, J. D., Rosen, H. J., Phengrasamy, L., & Miller, B. L. (2004). Cognitive and behavioral profile in a case of right anterior temporal lobe neurodegeneration. *Cortex*, *40*(4-5), 631-644.

Gough, P. M., Nobre, A. C., & Devlin, J. T. (2005). Dissociating linguistic processes in the left inferior frontal cortex with transcranial magnetic stimulation. *Journal of Neuroscience*, *25*(35), 8010-8016.

Griffiths, T. (2002). Gibbs sampling in the generative model of latent dirichlet allocation.

Griffiths, J.D., Marslen-Wilson, W.D., Stamatakis, E.A. & Tyler L.K. (2013). Functional Organization of the Neural Language System: Dorsal and Ventral Pathways Are Critical for Syntax. *Cerebral Cortex*, 23(1), 139-147.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(suppl 1), 5228-5235.

Grodzinsky, Y., & Friederici, A. D. (2006). Neuroimaging of syntax and syntactic processing. *Current Opinion in Neurobiology*, *16*(2), 240-246.

Gunter, T. C., Friederici, A. D., & Schriefers, H. (2000). Syntactic gender and semantic expectancy: ERPs reveal early autonomy and late interaction. *Journal of Cognitive Neuroscience*, *12*(4), 556-568.

Gunter, T. C., Stowe, L. A., & Mulder, G. (1997). When syntax meets semantics. *Psychophysiology*, *34*(6), 660-676.

Hahne, A., & Friederici, A. D. (1999). Electrophysiological evidence for two steps in syntactic analysis: Early automatic and late controlled processes. *Journal of Cognitive Neuroscience*, *11*(2), 194-205.

Hahne, A., & Jescheniak, J. D. (2001). What's left if the Jabberwock gets the semantics? An ERP investigation into semantic and syntactic processes during auditory sentence comprehension. *Cognitive Brain Research*, *11*(2), 199-212.

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). Association for Computational Linguistics.

Hale, J. (2006). Uncertainty about the rest of the sentence. *Cognitive Science*, *30*(4), 643-672.

Hämäläinen, M. S., & Ilmoniemi, R. J. (1994). Interpreting magnetic fields of the brain: minimum norm estimates. *Medical & Biological Engineering & Computing*, *32*(1), 35-42.

Hauk, O., & Pulvermüller, F. (2004). Effects of word length and frequency on the human event-related potential. *Clinical Neurophysiology*, *115*(5), 1090-1103.

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science*, *304*(5669), 438-441.

Hagoort, P. (2005). On Broca, brain, and binding: a new framework. *Trends in Cognitive Sciences*, *9*(9), 416-423.

Hagoort, P. (2013). MUC (memory, unification, control) and beyond. *Frontiers in Psychology*, *4*, 416.

Hare, M., McRae, K., & Elman, J. L. (2003). Sense and structure: Meaning as a determinant of verb subcategorization preferences. *Journal of Memory and Language*, *48*(2), 281-303.

Hare, M., McRae, K., & Elman, J. (2004). Admitting that admitting verb sense into corpus analyses makes sense. *Language and Cognitive Processes*, *19*(2), 181-224.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146-162.

Hartung, M., Kaupmann, F., Jebbara, S., & Cimiano, P. (2017). Learning Compositionality Functions on Word Embeddings for Modelling Attribute Meaning in Adjective-Noun Phrases. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers* (Vol. 1, pp. 54-64).

Hengeveld, K. (2004). The architecture of a functional discourse grammar. *A new architecture for Functional Grammar,* 1-21.

Hengeveld, K., & Mackenzie, J. L. (2008). *Functional Discourse Grammar: A typologically-based theory of language structure.* Oxford University Press.

Henson, R. N., Mouchlianitis, E., & Friston, K. J. (2009). MEG and EEG data fusion: simultaneous localisation of face-evoked responses. *Neuroimage*, *47*(2), 581-589.

Hickok, G., & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences*, *4*(4), 131-138.

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition*, *92*(1-2), 67-99.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393.

Hillebrand, A., & Barnes, G. R. (2002). A quantitative assessment of the sensitivity of whole-head MEG to activity in the adult human cortex. *Neuroimage*, *16*(3), 638-650.

Hochreiter, S., & Schmidhuber, J. (1997). LSTM can solve hard long time lag problems. *Advances in neural information processing systems,* 473-479.

Humphries, C., Binder, J. R., Medler, D. A., & Liebenthal, E. (2007). Time course of semantic processes during sentence comprehension: an fMRI study. *Neuroimage*, *36*(3), 924-932.

Jennings, F., Randall, B., & Tyler, L. K. (1997). Graded effects of verb subcategory preferences on parsing: Support for constraint-satisfaction models. *Language and Cognitive Processes*, *12*(4), 485-504.

Johnson-Laird, P. N., & Byrne, R. M. (2002). Conditionals: a theory of meaning, pragmatics, and inference. *Psychological Review*, *109*(4), 646.

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., & Wu, Y. (2016). Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410.*

Jung-Beeman, M. (2005). Bilateral brain processes for comprehending natural language. *Trends in Cognitive Sciences*, *9*(11), 512-518.

Just, M. A., Carpenter, P. A., Keller, T. A., Eddy, W. F., & Thulborn, K. R. (1996). Brain activation modulated by sentence comprehension. *Science*, *274*(5284), 114-116.

Kamide, Y., Altmann, G. T., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, *49*(1), 133-156.

Kan, I. P., & Thompson-Schill, S. L. (2004). Selection from perceptual and conceptual representations. *Cognitive, Affective, & Behavioral Neuroscience*, *4*(4), 466-482.

Kang, A. M., Constable, R. T., Gore, J. C., & Avrutin, S. (1999). An event-related fMRI study of implicit phrase-level syntactic and semantic processing. *Neuroimage*, *10*(5), 555-561.

Khateb, A., Pegna, A. J., Landis, T., Mouthon, M. S., & Annoni, J. M. (2010). On the origin of the N400 effects: an ERP waveform and source localization analysis in three matching tasks. *Brain Topography*, *23*(3), 311-320.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Skip-thought vectors. In *Advances in neural information processing systems* (pp. 3294-3302).

Klimovich-Gray, A., Tyler, L. K., Randall, B., Kocagoncu, E., Devereux, B., & Marslen-Wilson, W. D. (2019). Balancing Prediction and Sensory Input in Speech Comprehension: The Spatiotemporal Dynamics of Word Recognition in Context. *Journal of Neuroscience*, *39*(3), 519-527.

Kocagoncu, E., Clarke, A., Devereux, B. J., & Tyler, L. K. (2017). Decoding the cortical dynamics of sound-meaning mapping. *Journal of Neuroscience*, *37*(5), 1312-1319.

Korhonen, A., Krymolowski, Y., & Briscoe, T. (2006). A large subcategorization lexicon for natural language processing applications. In *Proceedings of LREC* (Vol. 6, pp. 1015-1020).

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4.

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*(8), 401-412.

Kutas, M. (1993). In the company of other words: Electrophysiological evidence for single-word and sentence context effects. *Language and Cognitive Processes*, *8*(4), 533-572.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual review of Psychology*, *62*, 621-647.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203-205.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension?. *Language, Cognition and Neuroscience*, *31*(1), 32-59.

Kuperberg, G. R. (2016). Separate streams or probabilistic inference? What the N400 can tell us about the comprehension of events. *Language, Cognition and Neuroscience*, *31*(5), 602-616.

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics:(de) constructing the N400. *Nature Reviews Neuroscience*, *9*(12), 920.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126-1177.

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *arXiv preprint arXiv:1611.01368*.

Liu, A. A., Su, Y. T., Jia, P. P., Gao, Z., Hao, T., & Yang, Z. X. (2015). Multiple/single-view human action recognition via part-induced multitask structural learning. *IEEE transactions on cybernetics*, *45*(6), 1194-1208.

López, J. D., Litvak, V., Espinosa, J. J., Friston, K., & Barnes, G. R. (2014). Algorithmic procedures for Bayesian MEG/EEG source reconstruction in SPM. *NeuroImage*, *84*, 476-487

Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology*, *88*, 22-60.

Luong, M. T., Le, Q. V., Sutskever, I., Vinyals, O., & Kaiser, L. (2015). Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*

MacDonald, M. C. (1994). Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes*, *9*(2), 157-201.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*(4), 676.

Maess, B., Mamashli, F., Obleser, J., Helle, L., & Friederici, A. D. (2016). Prediction signatures in the brain: semantic pre-activation during language comprehension. *Frontiers in Human Neuroscience*, *10*, 591.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177-190.

Marslen-Wilson, W. (1973). Linguistic structure and speech shadowing at very short latencies. *Nature*, *244*(5417), 522.

Marslen-Wilson, W. D. (1975). Sentence perception as an interactive parallel process. *Science*, *189*(4198), 226-228.

Marslen-Wilson, W., Brown, C. M., & Tyler, L. K. (1988). Lexical representations in spoken language comprehension. *Language and Cognitive Processes*, *3*(1), 1-16.

Marslen-Wilson, W., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*, *8*(1), 1-71.

Marslen-Wilson, W., & Tyler, L. K. (1987). Against modularity. *Modularity in knowledge representation and natural language understanding*, 37-62.

Marslen-Wilson, W. D., Tyler, L. K., & Koster, C. (1993). Integrative processes in utterance resolution. *Journal of Memory and Language*, *32*(5), 647-666.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, *10*(1), 29-63.

Martin, J. H., & Jurafsky, D. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River: Pearson/Prentice Hall.Jackendoff, R. (1997). *The architecture of the language faculty*(No. 28). MIT Press.

Mattout, J., Henson, R. N., & Friston, K. J. (2007). Canonical source reconstruction for MEG. *Computational Intelligence and Neuroscience*, *2007*.

Mazoyer, B. M., Tzourio, N., Frak, V., Syrota, A., Murayama, N., Levrier, O., ... & Mehler, J. (1993). The cortical representation of speech. *Journal of Cognitive Neuroscience*, *5*(4), 467-479.

McCarthy, D. (2001). *Lexical acquisition at the syntax-semantics interface: diathesis alternations, subcategorization frames and selectional preferences* (Doctoral dissertation, University of Sussex).

McRae, K., De Sa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, *126*(2), 99

McRae, K., Cree, G. S., Westmacott, R., & Sa, V. R. D. (1999). Further evidence for feature correlations in semantic memory. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *53*(4), 360.

Mellish, C. S. (1981). Coping with uncertainty: Noun phrase interpretation and early semantic analysis.

Mesgarani, N., David, S. V., Fritz, J. B., & Shamma, S. A. (2008). Phoneme representation and classification in primary auditory cortex. *The Journal of the Acoustical Society of America*, *123*(2), 899-909.

Metusalem, R., Kutas, M., Urbach, T. P., Hare, M., McRae, K., & Elman, J. L. (2012). Generalized event knowledge activation during online sentence comprehension. *Journal of Memory and Language*, *66*(4), 545-567.

Miglioretti, D. L., & Boatman, D. (2003). Modeling variability in cortical representations of human complex sound perception. *Experimental Brain Research*, *153*(3), 382-387.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39-41.

Minka, T. (2000). Estimating a Dirichlet distribution.

Monsalve, I. F., Frank, S. L., & Vigliocco, G. (2012, April). Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 398-408). Association for Computational Linguistics.

Morosan, P., Rademacher, J., Schleicher, A., Amunts, K., Schormann, T., & Zilles, K. (2001). Human primary auditory cortex: cytoarchitectonic subdivisions and mapping into a spatial reference system. *Neuroimage*, *13*(4), 684-701

Mummery, C. J., Patterson, K., Price, C. J., Ashburner, J., Frackowiak, R. S., & Hodges, J. R. (2000). A voxel-based morphometry study of semantic dementia: relationship between temporal lobe atrophy and semantic memory. *Annals of Neurology*, *47*(1), 36-45.

Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., ... & Allik, J. (1997). Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature*, *385*(6615), 432.

Narain, C., Scott, S. K., Wise, R. J., Rosen, S., Leff, A., Iversen, S. D., & Matthews, P. M. (2003). Defining a left-lateralized response specific to intelligible speech using fMRI. *Cerebral Cortex*, *13*(12), 1362-1368.

Ni, W., Constable, R. T., Mencl, W. E., Pugh, K. R., Fulbright, R. K., Shaywitz, S. E., ... & Shankweiler, D. (2000). An event-related neuroimaging study distinguishing form and content in sentence processing. *Journal of Cognitive Neuroscience*, *12*(1), 120-133.

Nieuwland, M. S., & Van Berkum, J. J. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience*, *18*(7), 1098-1111.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS Computational Biology*, *10*(4), e1003553.

O'Brien, G. (1998). Connectionism, analogicity and mental content. *Acta Analytica*, *22*, 111-131.

Obleser, J., & Kotz, S. A. (2009). Expectancy constraints in degraded speech modulate the language comprehension network. *Cerebral Cortex*, *20*(3), 633-640.

Ó Séaghdha, D., & Korhonen, A. (2014). Probabilistic distributional semantics with latent variable models. *Computational Linguistics*, *40*(3), 587-631.

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, *31*(6), 785-806.

Osterhout, L., Holcomb, P. J., & Swinney, D. A. (1994). Brain potentials elicited by garden-path sentences: evidence of the application of verb information during parsing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 786.

Peelle, J. E., Troiani, V., & Grossman, M. (2009). Interaction between process and content in semantic memory: An fMRI study of noun feature knowledge. *Neuropsychologia*, *47*(4), 995-1003.

Perani, D., Cappa, S. F., Schnur, T., Tettamanti, M., Collina, S., Rosa, M. M., & Fazio1, F. (1999). The neural correlates of verb and noun processing: A PET study. *Brain*, *122*(12), 2337-2344.

Pobric, G., Jefferies, E., & Ralph, M. A. L. (2007). Anterior temporal lobes mediate semantic representation: mimicking semantic dementia by using rTMS in normal participants. *Proceedings of the National Academy of Sciences*, *104*(50), 20137-20141.

Price, C. J. (2010). The anatomy of language: a review of 100 fMRI studies published in 2009. *Annals of the new York Academy of Sciences*, *1191*(1), 62-88.

Price, C. J. (2012). A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. *Neuroimage*, *62*(2), 816-847.

Rayner, K., Carlson, M., & Frazier, L. (1983). The interaction of syntax and semantics during sentence processing: Eye movements in the analysis of semantically biased sentences. *Journal of Memory and Language*, *22*(3), 358.

Rice, G. E., Lambon Ralph, M. A., & Hoffman, P. (2015). The roles of left versus right anterior temporal lobes in conceptual knowledge: an ALE meta-analysis of 97 functional neuroimaging studies. *Cerebral Cortex*, *25*(11), 4374-4391.

Roark, B., Bachrach, A., Cardenas, C., & Pallier, C. (2009, August). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1* (pp. 324-333). Association for Computational Linguistics.

Rodd, J. M., Davis, M. H., & Johnsrude, I. S. (2005). The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity. *Cerebral Cortex*, *15*(8), 1261-1269.

Rodd, J. M., Longe, O. A., Randall, B., & Tyler, L. K. (2010). The functional organisation of the fronto-temporal language system: evidence from syntactic and semantic ambiguity. *Neuropsychologia*, *48*(5), 1324-1335.

Röder, B., Stock, O., Neville, H., Bien, S., & Rösler, F. (2002). Brain activation modulated by the comprehension of normal and pseudo-word sentences of different processing demands: a functional magnetic resonance imaging study. *Neuroimage*, *15*(4), 1003-1014

Rogalsky, C., & Hickok, G. (2008). Selective attention to semantic and syntactic features modulates sentence processing networks in anterior temporal cortex. *Cerebral Cortex*, *19*(4), 786-796.

Rolheiser, T., Stamatakis, E. A., & Tyler, L. K. (2011). Dynamic processing in the human language system: synergy between the arcuate fascicle and extreme capsule. *Journal of Neuroscience*, *31*(47), 16949-16957.

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1987). Parallel Processing.

Sag, I. A., & Wasow, T. (2011). Performance-compatible competence grammar. *Non-transformational syntax: Formal and explicit models of grammar*, 359-377.

Sato, M., Yamashita, O., Sato, M. A., & Miyawaki, Y. (2018). Information spreading by a combination of MEG source estimation and multivariate pattern classification. *PloS one*, *13*(6), e0198806.

Saur, D., Kreher, B. W., Schnell, S., Kümmerer, D., Kellmeyer, P., Vry, M. S., ... & Huber, W. (2008). Ventral and dorsal pathways for language. *Proceedings of the National Academy of Sciences*, *105*(46), 18035-18040.

Schlösser, R., Hutchinson, M., Joseffer, S., Rusinek, H., Saarimaki, A., Stevenson, J., ... & Brodie, J. D. (1998). Functional magnetic resonance imaging of human brain activity in a verbal fluency task. *Journal of Neurology, Neurosurgery & Psychiatry*, *64*(4), 492-498.

Seidenberg, M. S., MacDonald, M. C., & Saffran, J. R. (2002). Does grammar start where statistics stop?. *Science*, *298*(5593), 553-554.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379-423.

Sharon, D., Hämäläinen, M. S., Tootell, R. B., Halgren, E., & Belliveau, J. W. (2007). The advantage of combining MEG and EEG: comparison to fMRI in focally stimulated visual cortex. *Neuroimage*, *36*(4), 1225-1235.

Simos, P. G., Basile, L. F., & Papanicolaou, A. C. (1997). Source localization of the N400 response in a sentence-reading paradigm using evoked magnetic fields and magnetic resonance imaging. *Brain Research*, *762*(1-2), 29-39.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302-319.

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage*, *44*(1), 83-98.

Snijders, T. M., Petersson, K. M., & Hagoort, P. (2010). Effective connectivity of cortical and subcortical regions during unification of sentence structure. *Neuroimage*, *52*(4), 1633-1644.

Snijders, T. M., Vosse, T., Kempen, G., Van Berkum, J. J., Petersson, K. M., & Hagoort, P. (2008). Retrieval and unification of syntactic structure in sentence comprehension: an fMRI study using word-category ambiguity. *Cerebral Cortex*, *19*(7), 1493-1503.

Sohoglu, E., & Davis, M. H. (2016). Perceptual learning of degraded speech by minimizing prediction error. *Proceedings of the National Academy of Sciences*, *113*(12), E1747-E1756.

Sohoglu, E., Peelle, J. E., Carlyon, R. P., & Davis, M. H. (2012). Predictive top-down integration of prior knowledge during speech perception. *Journal of Neuroscience*, *32*(25), 8443-8453.

. Sohoglu, E., Peelle, J. E., Carlyon, R. P., & Davis, M. H. (2014). Top-down influences of written text on perceived clarity of degraded speech. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(1), 186.

Sormaz, M., Murphy, C., Wang, H. T., Hymers, M., Karapanagiotidis, T., Poerio, G., ... & Smallwood, J. (2018). Default mode network can support the level of detail in experience during active task states. *Proceedings of the National Academy of Sciences*, *115*(37), 9318-9323.

Spivey-Knowlton, M. J., Trueswell, J. C., & Tanenhaus, M. K. (1993). Context effects in syntactic ambiguity resolution: Discourse and semantic influences in parsing reduced relative clauses. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *47*(2), 276.

Steinhauer, K., & Drury, J. E. (2012). On the early left-anterior negativity (ELAN) in syntax studies. *Brain and Language*, *120*(2), 135-162.

Su, L., Fonteneau, E., Marslen-Wilson, W., & Kriegeskorte, N. (2012). Spatiotemporal searchlight representational similarity analysis in EMEG source space. In *Pattern recognition in neuroimaging (prni), 2012 international workshop on* (pp. 97-100). IEEE.

Sundermeyer, M., Ney, H., & Schlüter, R. (2015). From feedforward to recurrent LSTM neural networks for language modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(3), 517-529.

Tanabe, H., Sawada, T., Inoue, N., Ogawa, M., Kuriyama, Y., & Shiraishi, J. (1987). Conduction aphasia and arcuate fasciculus. *Acta Neurologica Scandinavica*, *76*(6), 422-427.

Taraban, R., & McClelland, J. L. (1988). Constituent attachment and thematic role assignment in sentence processing: Influences of content-based expectations. *Journal of Memory and Language*, *27*(6), 597-632.

Taulu, S., Simola, J., & Kajola, M. (2005). Applications of the signal space separation method. *IEEE transactions on signal processing*, *53*(9), 3359-3372.

Taylor, W. L. (1953). “Cloze procedure”: A new tool for measuring readability. *Journalism Bulletin*, *30*(4), 415-433.

Thompson, H. E., Henshall, L., & Jefferies, E. (2016). The role of the right hemisphere in semantic control: A case-series comparison of right and left hemisphere stroke. *Neuropsychologia*, *85*, 44-61.

Trueswell, J. C., Tanenhaus, M. K., & Kello, C. (1993). Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(3), 528.

Trueswell, J. C., Tanenhaus, M. K., & Garnsey, S. M. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*, *33*, 285-285.

Tyler, L. K., Cheung, T. P., Devereux, B. J., & Clarke, A. (2013). Syntactic computations in the language network: characterizing dynamic network properties using representational similarity analysis. *Frontiers in Psychology*, *4*, 271.

Tyler, L. K., & Marslen-Wilson, W. D. (1977). The on-line effects of semantic context on syntactic processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(6), 683-692.

Tyler, L. K., & Marslen-Wilson, W. (2008). Fronto-temporal brain systems supporting spoken language comprehension. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 1037-1054.

Tyler, L. K., Stamatakis, E. A., Post, B., Randall, B., & Marslen-Wilson, W. (2005). Temporal and frontal systems in speech comprehension: An fMRI study of past tense processing. *Neuropsychologia*, *43*(13), 1963-1974.

Tyler, L.K., Wright, P., Randall, B., Marslen-Wilson, W.D., & Stamatakis, E.A. (2010). Reorganisation of syntactic processing following LH brain damage: Does RH activity preserve function? *Brain, 133(11),*3396-3408.

Van Petten, C., & Kutas, M. (1990). Interactions between sentence context and word frequencyinevent-related brainpotentials. *Memory & Cognition*, *18*(4), 380-393.

Vartanian, O., & Goel, V. (2005). Task constraints modulate activation in right ventral lateral prefrontal cortex. *Neuroimage*, *27*(4), 927-933.

Vigneau, M., Beaucousin, V., Herve, P. Y., Duffau, H., Crivello, F., Houde, O., ... & Tzourio-Mazoyer, N. (2006). Meta-analyzing left hemisphere language areas: phonology, semantics, and sentence processing. *Neuroimage*, *30*(4), 1414-1432.

Wallach, H. M. (2002). Structured topic models for language. Unpublished doctoral dissertation, University of Cambridge

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

Wicha, N. Y., Moreno, E. M., & Kutas, M. (2004). Anticipating words and their gender: An event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *Journal of Cognitive Neuroscience*, *16*(7), 1272-1288.

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & Van den Bosch, A. (2015). Prediction during natural language comprehension. *Cerebral Cortex*, *26*(6), 2506-2516.

Xu, J., Kemeny, S., Park, G., Frattali, C., & Braun, A. (2005). Language in context: emergent features of word, sentence, and narrative comprehension. *Neuroimage*, *25*(3), 1002-1015.

Ye, Z., & Zhou, X. (2009). Conflict control during sentence comprehension: fMRI evidence. *Neuroimage*, *48*(1), 280-290.

Zhu, X., Sobihani, P., & Guo, H. (2015, June). Long short-term memory over recursive structures. In *International Conference on Machine Learning* (pp. 1604-1612).

Zhuang, J., Tyler, L. K., Randall, B., Stamatakis, E. A., & Marslen-Wilson, W. D. (2012). Optimally efficient neural systems for processing spoken language. *Cerebral Cortex*, *24*(4), 908-918.

Appendices

Appendix 1: Merge and integration

One of the most important operations in language understanding is to combine the aforementioned (lexically activated) information to sketch a comprehensive picture of the intended message. Although this claim is widely acknowledged in linguistics, the way in which such combinatorial operation (or integration) occurs is still controversial. In the light of Chomsky's minimalist program (1993), "merge" is described as one of the basic phrase structure operations that combine two syntactic objects at the root to form a new object, inhibiting the features of an object that are incompatible with its sister; for example, after merging "kick" and "a ball" into a verb phrase, the features of "kick" as a noun will be inhibited (if all of the features are incompatible, the sentence is not grammatical). Here, syntactic objects refer to the nodes in a syntactic tree diagram from lexical to phrasal or clausal items. This merge process is recursive: it combines the syntactic objects at the root and this newly combined object is then combined with its sister and so on until it reaches a maximal projection of the tree. The maximal projection refers to a node that cannot be projected further and, in this recursive paradigm, the maximal projection of two objects becomes an intermediate projection at the later stage when combining it with its sister. Hence, this entire processing scheme is bottom-up driven, based on the binary branching (hence, consistent with the *x-bar theory*[1]) and constituency-based phrase structure grammar (as opposed to dependency grammar). This theory of merge is rejected by many other grammar theories including the LFG and dependency grammar due to these assumptions. In an interactive view that describes the human language system as a predictive machine, more plausible models of such combinatorial processing must incorporate the top-down influence on processing the bottom-up input. In a recent generative probabilistic model of human language processing (Kuperberg, 2016; Kuperberg & Jaeger, 2016), integration refers to the process of adapting the system's beliefs with respect to the bottom-up input at a number of different linguistic levels (see Chapter 2 for more details).

[1] *The x-bar theory (Chomsky, 1970; Jackendoff, 1977) describes the internal structure of constituents or syntactic objects based on the notion that all phrases share some essential structural properties. It is basically a template that reduces all phrase structures (XP) to*

recursive specifier-head configurations with x-bar (denoted as X’) being an intermediate projection of the head (X). In this theory, X refers to any arbitrary lexical category which, in real practice, is often replaced with V for a verb, A for an adjective, N for a noun or P for a preposition. The constraining rules of phrase structure grammar are its central properties which includes; 1. An X-phrase consists of an optional specifier and an X’, 2. An X’ could dominate another X’ and an adjunct and 3. A head X and its complement are sisters dominated by their mother X’. Note that the concept of “projection” originated from this x-bar theory defined as any X^N being a projection of X^0 (N (number of bars) > 0). In practice, various functions can be assigned to the specifier position depending on the category of X (or maximal projection of X): for example, it could be a determiner of NP (e.g. ‘a’ or ‘the’), a degree element of AP (e.g. ‘few’, ‘several’, ‘some’, ‘many’ etc.), subject of IP (see figure below) or a modifier of VP (e.g. adverb). The figure below illustrates syntactic parsing of an example sentence “The experienced walker chose the path that ran by the river” based on this x-bar theory.

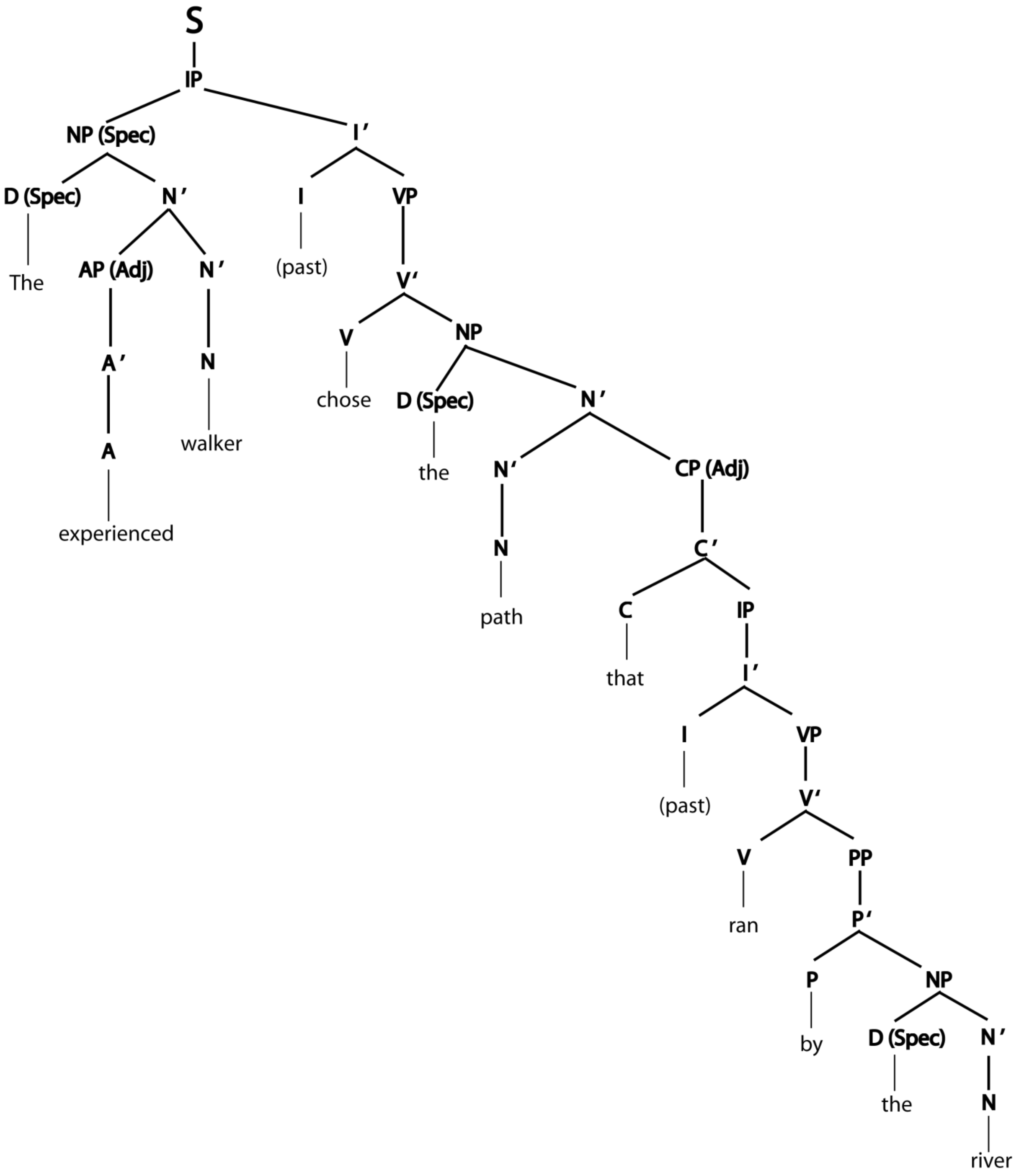


Figure 2: A visual illustration of the x-bar parsing of a sentence. Note that the specifiers and adjuncts are highlighted by (Spec) and (Adj). Abbreviations: IP = inflectional phrase, NP = noun phrase, D = determiner, AP = adjectival phrase, I = inflection, VP = verb phrase, V = verb, CP = complement phrase, C = complement word, PP = prepositional phrase and P = preposition.

A1-References

Chomsky, N (1970). Remarks on nominalization. In: R. Jacobs and P. Rosenbaum (eds.) *Reading in English Transformational Grammar,* 184-221. Waltham: Ginn

Chomsky, N (1993). *A minimalist program for linguistic theory*. MIT occasional papers in linguistics no. 1. Cambridge, MA: Distributed by MIT Working Papers in Linguistics

Jackendoff, R (1977). *X-bar-Syntax: A Study of Phrase Structure*. Linguistic Inquiry Monograph 2. Cambridge, MA:MIT Press

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension?. *Language, cognition and neuroscience*, *31*(1), 32-59.

Kuperberg, G. R. (2016). Separate streams or probabilistic inference? What the N400 can tell us about the comprehension of events. *Language, cognition and neuroscience*, *31*(5), 602-616.

Appendix 2: A list of all sentence stimuli

Table A2: all sentence stimuli are shown in conjunction with the experimental manipulation. The column specified as "SCFsurp" shows whether a particular syntactic frame of the complement in each sentence is more (low surprisal) or less expected (high surprisal) given a preceding verb. This SCF surprisal was computed using VALEX database. Similarly, the column titled as "Argsurp" shows whether a particular content word in the complement is more (low surprisal) or less expected (high surprisal). This argument surprisal was computed from the Google Ngram database (https://books.google.com/ngrams)

Sentences	SCFsurp	Argsurp
The bank manager acknowledged the difference between the two sums	low	low
The bank manager acknowledged the leader of the campaigning group	low	high
The bank manager acknowledged that the decision had been made quickly	high	low
The bank manager acknowledged that the argument had been heated	high	high
The clever man adapted to the role of house husband	low	low
The clever man adapted to the community in the remote town	low	high
The clever man adapted the play for the silver screen	high	low
The clever man adapted the hospital for disabled people	high	high
The proud woman announced the birth of her first grandchild	low	low
The proud woman announced the progress of the fundraising project	low	high
The proud woman announced that the sale had raised a million pounds	high	low
The proud woman announced that the appeal had exceeded its target	high	high
The graduate student applied for the post of part-time lecturer	low	low
The graduate student applied for the test to be delayed	low	high
The graduate student applied the technique to his research	high	low
The graduate student applied the skill to designing experiments	high	high
The elderly lady appreciated the help from her next door neighbours	low	low
The elderly lady appreciated the dog who had been her pet for years	low	high
The elderly lady appreciated that the purpose of the visit was good	high	low
The elderly lady appreciated that the support would end in December	high	high
The busy secretary arranged the ceremony to welcome her new boss	low	low
The busy secretary arranged the clothing that was hanging in the wardrobe	low	high
The busy secretary arranged for the publication of the latest accounts	high	low
The busy secretary arranged for the approval of the cleaning contract	high	high
The brave firefighters attempted to cope with the leaping flames	low	low
The brave firefighters attempted to warn people to stay away	low	high
The brave firefighters attempted the search in difficult circumstances	high	low
The brave firefighters attempted the procedure to save the man's life	high	high
The police officer believed the story about the hidden gun	low	low
The police officer believed the result of the investigation	low	high
The police officer believed that the death was extremely suspicious	high	low
The police officer believed that the evening was when criminals struck	high	high
The experienced walker chose the path that ran by the river	low	low
The experienced walker chose the card to send to his mother	low	high
The experienced walker chose to abandon his rucksack by the hedge	high	low

The experienced walker chose to relax with his feet in the stream	high	high
The naughty child climbed on the back of his grandmother's chair	low	low
The naughty child climbed on the top of the kitchen cupboard	low	high
The naughty child climbed the tree at the bottom of the garden	high	low
The naughty child climbed the bank to get his football back	high	high
The duty solicitor concluded that the election had been fixed after all	low	low
The duty solicitor concluded that the lunch was the best he had tasted	low	high
The duty solicitor concluded the discussion of his client's case	high	low
The duty solicitor concluded the battle to access his client's records	high	high
The elderly couple continued to travel around town by bus	low	low
The elderly couple continued to thank their daughter for her help	low	high
The elderly couple continued the conversation about the war	high	low
The elderly couple continued the holiday in spite of their colds	high	high
The TV announcer declared the death of the president with sadness	low	low
The TV announcer declared the result of the election at noon	low	high
The TV announcer declared that the law had been passed	high	low
The TV announcer declared that the road would be closed from midnight	high	high
The timid man declined to share the results of the survey wth his friends	low	low
The timid man declined to touch the slimy mixture in the bowl	low	high
The timid man declined the opportunity to meet the famous film star	high	low
The timid man declined the drink that contained lots of alcohol	high	high
The accused man denied the benefit of having a defence lawyer	low	low
The accused man denied the evidence of the police officer	low	high
The accused man denied that the court had the right to try him	high	low
The accused man denied that the trouble was caused by his drinking	high	high
The diligent headteacher deserved the attention she got from the parents	low	low
The diligent headteacher deserved the deal she made about her salary	low	high
The diligent headteacher deserved to win praise from all the staff	high	low
The diligent headteacher deserved to arrive late from time to time	high	high
The local politician emphasised the point of lowering speed on local roads	low	low
The local politician emphasised the system for claiming housing benefits	low	high
The local politician emphasised that the question must be debated	high	low
The local politician emphasised that the night would be great fun	high	high
The story writer engaged in the debate raging on the internet	low	low
The story writer engaged in the session about the use of imagery	low	high
The story writer engaged the imagination of many small children	high	low
The story writer engaged the editor in a long correspondence	high	high
The intrepid child found the picture before everyone else	low	low
The intrepid child found the teacher hiding in the staffroom	low	high
The intrepid child found that the activity made him hungry	high	low
The intrepid child found that the doubt made him hesitate	high	high
The young man fled the scene of the terrible accident	low	low
The young man fled the army when the fighting began	low	high
The young man fled to the forest when the chase began	high	low
The young man fled to the security of his friend's house	high	high
The absentminded professor forgot the promise he'd made to his student	low	low

The absentminded professor forgot the gap between the train and the platform	low	high
The absentminded professor forgot to inform his college that he was away	high	low
The absentminded professor forgot to boil his egg for four minutes	high	high
The walking couple heard the bird before they saw it	low	low
The walking couple heard the stone as it dropped into the water	low	high
The walking couple heard that the earth was completely waterlogged	high	low
The walking couple heard that the farm was open to visitors	high	high
The new worker helped the development with his carpentry skills	low	low
The new worker helped the window open with his elbow	low	high
The new worker helped to explain the plans to the residents	high	low
The new worker helped to catch the mouse in the office	high	high
The romantic student loved the snow on the college lawn	low	low
The romantic student loved the bridge near the city centre	low	high
The romantic student loved to dance at the college ball	high	low
The romantic student loved to jump into the sea at dawn	high	high
The assistant director managed to produce his action plan on time	low	low
The assistant director managed to wear a tie in the office	low	high
The assistant director managed the business for 25 years	high	low
The assistant director managed the effect of reduced staffing levels	high	high
The local vicar mentioned the name of the new curate in passing	low	low
The local vicar mentioned the street where the accident had happened	low	high
The local vicar mentioned that the word was mightier than the sword	high	low
The local vicar mentioned that the boy was singing in the choir	high	high
The determined father moved to the side of the room where his son stood	low	low
The determined father moved to the group that was causing the trouble	low	high
The determined father moved the family into a lovely brick house	high	low
The determined father moved the case to the middle of the platform	high	high
The stranded householder needed the aid that the Red Cross was sending	low	low
The stranded householder needed the discovery of a good escape route	low	high
The stranded householder needed to complete the repairs to his battered car	high	low
The stranded householder needed to dig the snow away from the front door	high	high
The factory manager neglected the potential of the new technology	low	low
The factory manager neglected the appointment with his best customer	low	high
The factory manager neglected to secure the doors yesterday evening	high	low
The factory manager neglected to display the health and safety rules	high	high
The nursery teacher planned the event at the primary school	low	low
The nursery teacher planned the music for the nativity play	low	high
The nursery teacher planned to sell some toys at the market	high	low
The nursery teacher planned to feed the hamster before lunchtime	high	high
The aid worker pleaded for the freedom to treat the injured soldiers	low	low
The aid worker pleaded for the care to be extended to boy	low	high
The aid worker pleaded the cause of sick children everywhere	high	low
The aid worker pleaded the condition that she leave by midnight	high	high
The football fans predicted the growth in penalty shoot outs	low	low
The football fans predicted the price of pies at the stadium	low	high
The football fans predicted that the future would bring many victories	high	low

The football fans predicted that the wind would blow the ball away	high	high
The unhappy driver preferred to listen to music in his car	low	low
The unhappy driver preferred to cause maximum trouble on the road	low	high
The unhappy driver preferred the chance of avoiding a fine	high	low
The unhappy driver preferred the doctor who never challenged him	high	high
The busy father prepared the meal for his children in the evening	low	low
The busy father prepared the response to his son's demands	low	high
The busy father prepared to claim a refund on his parking permit	high	low
The busy father prepared to survive his son's teenage years	high	high
The office manager promised the position to the best candidate	low	low
The office manager promised the table to the new recruit	low	high
The office manager promised to consider rewriting the report	high	low
The office manager promised to add typing to the job description	high	high
The rural residents protested the action taken by the local farmer	low	low
The rural residents protested the control exerted by the government	low	high
The rural residents protested against the use of chemicals locally	high	low
The rural residents protested against the policy of culling badgers	high	high
The eager technician realised that the disease might infect newborn babies	low	low
The eager technician realised that the computer dominated his life	low	high
The eager technician realised the possibility of inventing new equipment	high	low
The eager technician realised the advantage of getting to work early	high	high
The senior nurse recognised the family of the elderly patient	low	low
The senior nurse recognised the end of traditional healthcare	low	high
The senior nurse recognised that the government had supported hospitals	high	low
The senior nurse recognised that the money had been spent on drugs	high	high
The private investigators recovered the goods for the owners of the house	low	low
The private investigators recovered the cash from the supermarket robbery	low	high
The private investigators recovered from the shock of solving the crime	high	low
The private investigators recovered from the conflict between the drugs barons	high	high
The obstinate child refused to betray his classmates to the teacher	low	low
The obstinate child refused to spell any of the words correctly	low	high
The obstinate child refused the invitation from the headteacher	high	low
The obstinate child refused the pencil offered by his friend	high	high
The astounded woman remembered the dream that had troubled her in the night	low	low
The astounded woman remembered the artist from before he was famous	low	high
The astounded woman remembered that the solution involved lots of deception	high	low
The astounded woman remembered that the actor had several oscars	high	high
The Essex police searched for the name in the database	low	low
The Essex police searched for the reason behind the crimes	low	high
The Essex police searched the area for the little girl	high	low
The Essex police searched the home for any signs of drugs	high	high
The young couple settled on the hill with the pretty houses	low	low
The young couple settled on the film starring Clint Eastwood	low	high
The young couple settled the issue between themselves	high	low

The young couple settled the account at the local shop	high	high
The boy's mother started the engine before wiping the windscreen	low	low
The boy's mother started the diet at the beginning of April	low	high
The boy's mother started to record the funny things he said	high	low
The boy's mother started to vary what she gave him for breakfast	high	high
The junior barrister submitted the report just before the deadline	low	low
The junior barrister submitted the material for the judge to assess	low	high
The junior barrister submitted to the authority of the expert	high	low
The junior barrister submitted to the terms of the judge's ruling	high	high
The desparate family suffered the pain of losing their home	low	low
The desparate family suffered the danger of being evicted	low	high
The desparate family suffered from the lack of decent housing	high	low
The desparate family suffered from the threat of court action	high	high
The evil dictator threatened the peace of the whole continent	low	low
The evil dictator threatened the agreement with neighbouring countries	low	high
The evil dictator threatened to attack the freedom of the press	high	low
The evil dictator threatened to ignore the rulings of the court	high	high
The excited child tried to speak but the words stuck in her throat	low	low
The excited child tried to believe that Santa would bring his presents	low	high
The excited child tried the door to see if it would open	high	low
The excited child tried the book she had found in the library	high	high
The senior administrator understood the business of health care	low	low
The senior administrator understood the example of his boss	low	high
The senior administrator understood that the road would be repaired	high	low
The senior administrator understood that the window would never open	high	high
The young woman wanted to escape from her boring parents	low	low
The young woman wanted to collect lots of diamond rings	low	high
The young woman wanted the coat that was on sale in Harrods	high	low
The young woman wanted the career of a supermodel	high	high

Appendix 3: Weber-Fechner's Law

In psychology, surprisal has an appealing trait that it relates the objective prediction probability to the subjective error response via logarithm. In fact, logarithm is widely acknowledged as an accurate estimate of the psychophysical function, mapping the objective stimulus in the physical space onto the perceived experience in the psychological space in humans. Tracing back to 1860s, Gustav Fechner suggested that the perceived sensation is logarithmically related to the actual stimulus intensity in humans. The explicit formulation of this notion is derived from Weber's law stating that the smallest detectable increment (or JND = just noticeable difference) in the actual stimulus intensity is proportional to the initial intensity of it (e.g. adding a 0.5kg weight when holding a 5kg weight can easily be noticed compared to adding a 0.5kg weight on top of a 10kg weight). It is expressed as:

$$\Delta I = KI \ldots (A3.1)$$

where ΔI is the smallest increment (e.g. 0.5kg), I is the initial weight (e.g. 5kg weight) and K is some constant of their ratio (e.g. 0.1kg). Then, Fechner additionally defined a psychophysical function that translates this constant into the smallest increment in the psychological space:

$$\Delta P = c\frac{\Delta I}{I} \ldots (A3.2)$$

where c is some transition constant. To obtain the perceived stimulus intensity P, we simply integrate (A3.2):

$$P = c \log I + C \ldots (A3.3)$$

At some threshold of the stimulus intensity I_p, the perceived intensity becomes zero. Hence, the constant C can be expressed as a function of this threshold $C = -c \log I_p$ (solving for 70 after substituting $I \rightarrow I_p$). By substituting $C = -c \log I_p$, we obtain:

$$P = c \log \frac{I}{I_p} \ldots (A3.4)$$

This is known as Fechner's law describing the subjective experience of the stimulus intensity P as a logarithm of the objective intensity from a measurement device I. In our settings, modelling the prediction error using the surprisal metric translates the objective (physical) prediction to the subjective (psychological) perception of the error by using logarithm as the psychophysical mapping function.

Appendix 4: derivation of the LDA training algorithm (collapsed Gibbs sampler)

Gibbs sampling is a widely used training algorithm for Bayesian models which obtains a sequence of observations approximated from a specified distribution since direct sampling is difficult. The specified distribution is often randomly initialized in the beginning and constantly updated during training. An application of this method to LDA model training is described in Griffiths (2002); see also, Griffiths & Steyvers (2004); Wallach (2002); O'Seaghdha & Korhonen (2014). In contrast to the Variational Bayesian algorithm, this Gibbs sampling method does not assume independence among the model parameters and the latent variable. Hence, this approach leads to more accurate results when they are not independent in exchange for slow convergence. Given that the training samples were selected from a subset of corpus data constrained to be in a direct object frame, this method was used for training the model.

The central idea of this training algorithm is that, for i^th^ observation in the corpus, it assigns the value for the latent variablez_i, conditionally on the currently observed variable c_i and w_i as well as the latent variable values for all other observations z_{-i} such that:

$$P(z_i = j|z_{-i}, c_i, w) \propto P(w_i|z_i = j, z_{-i}, w_{-i})P(z_i = j|z_{-i}, c_i) \dots (A4.1)$$

Now, the question reduces to finding the word-topic term $P(w_i|z_i = j, z_{-i}, w_{-i})$ and the topic-document term$P(z_i = j|z_{-i}, c_i)$. First, we could write these terms in a form:

$$P(w_i|z_i = j, z_{-i}, w_{-i}) = \int P\left(w_i\middle|z_i = j, \emptyset_j\right)P\left(\emptyset_j\middle|z_{-i}, w_{-i}\right)d\emptyset_j \dots (A4.2)$$

$$P(z_i = j|z_{-i}, c_i) = \int P\left(z_i = j\middle|\theta_{c_i}\right)P(\theta_{c_i}|z_{-i})\, d\theta_{c_i} \dots (A4.3)$$

where $\emptyset_j$is a parameter with the multinomial distribution over words associated with j^th^ topic and θ_{c_i} is another parameter with the multinomial distribution over topics associated with a particular document c_i. Note that all other observations denoted by the subscript $-i$ become conditionally independent of the current observation denoted by the subscript i once these multinomial parameters (informing the distributions from which the topic associated with the current observation is sampled) are known.

This approach is called "collapsed" Gibbs sampler since it marginalizes these parameters. Given these parameters, the first terms in the integral of (A4.2) and (A4.3) are represented by: $\emptyset_{j,w_i} = P\left(w_i\middle|z_i = j, \emptyset_j\right)$ and $\theta_{c_i,j} = P\left(z_i = j\middle|\theta_{c_i}\right)$ respectively. The second terms in the integral of (A4.2) and (A4.3) are the posteriors of the parameters $\emptyset_j$ and θ_{c_i}which are, in turn, expressed as: $P\left(\emptyset_j\middle|z_{-i}, w_{-i}\right) \propto P\left(w_{-i}\middle|\emptyset_j, z_{-i}\right)P\left(\emptyset_j\right)$ (the involvement of z_{-i} term partitions the words into sets assigned to different topics so that only those assigned to topic j can influence $\emptyset_j$) and $P(\theta_{c_i}|z_{-i}) \propto P\left(z_{-i}\middle|\theta_{c_i}\right)P\left(\theta_{c_i}\right)$ where $P\left(\emptyset_j\right)$ and $P\left(\theta_{c_i}\right)$ are Dirichlet priors hyperparametrized by β and α respectively.

Combining these, (A4.2) and (A4.3) can be rewritten as the expected posterior of these parameters:

$$E_{pos}\left[\emptyset_{j,w_i}\right] = P(w_i|z_i = j, z_{-i}, w_{-i}) \propto \int \emptyset_{j,w_i} P\left(w_{-i}\middle|\emptyset_j, z_{-i}\right)P\left(\emptyset_j\right)d\emptyset_j \ldots (A4.4)$$

$$E_{pos}\left[\theta_{c_i,j}\right] = P(z_i = j|z_{-i}, c_i) \propto \int \theta_{c_i,j} P\left(z_{-i}\middle|\theta_{c_i}\right)P\left(\theta_{c_i}\right)d\theta_{c_i} \ldots (A4.5)$$

These terms can be expressed as:

$$E_{pos}\left[\emptyset_{j,w_i}\right] = \int \emptyset_{j,w_i} P\left(\emptyset_j\middle|z_{-i}, w_{-i}\right)d\emptyset_j = \int \emptyset_{j,w_i} \frac{\Gamma\left(\sum_v f_{-i,j}^{(v)} + \beta\right)}{\prod_v \Gamma\left(f_{-i,j}^{(v)} + \beta\right)} \prod_v \emptyset_{j,v}^{f_{-i,j}^{(v)}+\beta-1} d\emptyset_j$$

$$= \int \frac{\Gamma\left(\sum_v f_{-i,j}^{(v)} + \beta\right)}{\prod_v \Gamma\left(f_{-i,j}^{(v)} + \beta\right)} \emptyset_{j,w_i}^{f_{-i,j}^{(w_i)}+\beta-1+1} \prod_{v\neq w_i} \emptyset_{j,v}^{f_{-i,j}^{(v)}+\beta-1} d\emptyset_j \ldots (A4.6)$$

$$E_{pos}\left[\theta_{c_i,j}\right] = \int \theta_{c_i,j} P(\theta_{c_i}|z_{-i})\, d\theta_{c_i}$$

$$= \int \frac{\Gamma\left(\sum_j f_{-i,j}^{(c_i)} + \beta\right)}{\prod_j \Gamma\left(f_{-i,j}^{(c_i)} + \beta\right)} \theta_{c_i,z_i}^{f_{-i,z_i}^{(c_i)}+\alpha_{z_i}-1+1} \prod_{j\neq z_i} \theta_{c_i,j}^{f_{-i,j}^{(c_i)}+\alpha_j-1} d\theta_{c_i} \ldots (A4.7)$$

By setting $g_{-i,j}^{(v)} = f_{-i,j}^{(v)} + \beta \;\forall\, v \neq w_i$ and $g_{-i,j}^{(w_i)} = f_{-i,j}^{(w_i)} + \beta + 1$, we can express $g_{-i,j} = 1 + \sum_v f_{-i,j}^{(v)} + \beta$ where $w_i \in v$. Using a property of the gamma function that $\Gamma(a+1) = a\Gamma(a)$, following expressions can be derived: $\Gamma\left(g_{-i,j}^{(w_i)}\right) = \left(f_{-i,j}^{(w_i)} + \beta\right)\Gamma\left(f_{-i,j}^{(w_i)} + \beta\right)$ and $\Gamma\left(g_{-i,j}\right) = \left(\sum_v f_{-i,j}^{(v)} + \beta\right)\Gamma\left(\sum_v f_{-i,j}^{(v)} + \beta\right)$. By substituting these to (A4.6), we obtain:

$$E_{pos}\left[\emptyset_{j,w_i}\right] = \int \frac{\dfrac{\Gamma\left(g_{-i,j}\right)}{\left(\sum_v f_{-i,j}^{(v)} + \beta\right)}}{\dfrac{\Gamma\left(g_{-i,j}^{(w_i)}\right)}{\left(f_{-i,j}^{(w_i)} + \beta\right)} \prod_{v\neq w_i} \Gamma\left(g_{-i,j}^{(v)}\right)} \emptyset_{j,w_i}^{g_{-i,j}^{(w_i)}-1} \prod_{v\neq w_i} \emptyset_{j,w_i}^{g_{-i,j}^{(v)}-1} d\emptyset_j$$

$$= \frac{\left(f_{-i,j}^{(w_i)} + \beta\right)}{\left(\sum_v f_{-i,j}^{(v)} + \beta\right)} \int \frac{\Gamma\left(g_{-i,j}\right)}{\prod_v \Gamma\left(g_{-i,j}^{(v)}\right)} \prod_v \emptyset_{j,w_i}^{g_{-i,j}^{(v)}-1} d\emptyset_j$$

$$= \frac{\left(f_{-i,j}^{(w_i)} + \beta\right)}{\left(\sum_v f_{-i,j}^{(v)} + \beta\right)} \int P\left(\emptyset_j\middle|g\right)d\emptyset_j \ldots (A4.8)$$

Given the probability axiom that $\int P(\emptyset_j|g)d\emptyset_j = 1$, the expected posterior can be summarized as:

$$E_{pos}[\emptyset_{j,w_i}] = P(w_i|z_i = j, z_{-i}, w_{-i}) = \frac{f_{-i,j}^{(w_i)} + \beta}{f_{-i,j} + |W|\beta} \dots (A4.9)$$

Same logic can be applied to compute the document-topic parameter as below given the asymmetric hyperparameter α, recommended by Wallach et al. (2009):

$$E_{pos}[\theta_{c_i,j}] = P(z_i = j|z_{-i}, c_i) = \frac{f_{-i,j}^{(c_i)} + \alpha_j}{\sum_j f_{-i,j}^{(c_i)} + \alpha_j} \dots (A4.10)$$

where f represents the frequency count (i.e. $f_{-i,j}^{(w_i)}$ is the frequency of a word at the current observation i associated with a topic j after taking out a topic assignment at i; $f_{-i,j}^{(c_i)}$ is the frequency of a topic j associated with the document c at the current observation i after taking out a topic assignment at i), $|W|$ represents the word-vector length (i.e. total number of words in the vocabulary) and $f_j = \sum_{w_i} f_j^{(w_i)}$.

One of the main advantages of this approach over VB is its less biased estimate. Given that the parameters are randomly initialized in the beginning, it is necessary to wait until the sampler "settles down" (this period of waiting is known as "burn-in" period). To improve the predictive stability, I averaged the document-topic (DT; see (A4.10)) and topic-word (TW; see (A4.9)) distributions computed from three independent sampling states (i.e. there was 50 iterations gap between each of these three sampling states to prevent auto-correlation) after the burn-in period of 200. These details were followed from O'Seaghdha and Korhonen (2014). To preserve the fine-grained pattern across topics while preventing redundancy, I set the total number of topics to 100.

The last remaining question is how to set values for the hyper-parameters α and β. The underlying notion of maximising the Dirichlet likelihood with respect to a parameter α is based on the fact that the Dirichlet is a member of the exponential family such that it could be written in a form:

$$P(x|\eta) = h(x)\exp\{\eta^T T(x) - A(\eta)\} \dots (A4.11)$$

with the following specifications: $h(x) = 1, \eta = \alpha - 1, T(x) = \log P$ and $A(\eta) = N(\sum_k \log \Gamma(\alpha_k) - \log \Gamma(\sum_k \alpha_k))$. Here, $A(\eta)$ is a convex function known as the cumulant generating function and, consequently, the log-likelihood of the data based on this function is also convex in η (and α) which guarantees a unique optimum:

$$\log P(x|\eta) = N\sum_{k}(\alpha_k - 1)\log P_k + N\log \Gamma\left(\sum_{k}\alpha_k\right) - N\sum_{k}\log \Gamma(\alpha_k) \ldots (A4.12)$$

where $P_k = \frac{1}{N}\sum_i \log P_{ik}$ for every data sample i. However, our objective function (evidence) follows the compound Dirichlet-multinomial distribution and, as with the Dirichlet likelihood, it does not have a closed-form solution. My optimisation procedure strictly follows Thomas Minka's fixed-point iteration scheme (Minka, 2000) which computes the lower-bound, convex in and tight at α, based on the initial guess of α. Using the maximum of this bound in closed-form as a new guess, the optimisation scheme iterates until convergence.

A4-references

Griffiths, T. (2002). Gibbs sampling in the generative model of latent dirichlet allocation.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, *101*(suppl 1), 5228-5235.

Minka, T. (2000). Estimating a Dirichlet distribution.

Ó Séaghdha, D., & Korhonen, A. (2014). Probabilistic distributional semantics with latent variable models. *Computational Linguistics*, *40*(3), 587-631.

Wallach, H. M. (2002). Structured topic models for language. Unpublished doctoral dissertation, University of Cambridge

Appendix 5: Representing the constraint of each verb on its argument through an optimized set of "synsets" (i.e. conceptual senses)

Tuning the WordNet conceptual hierarchy using the VALEX lexical constraint

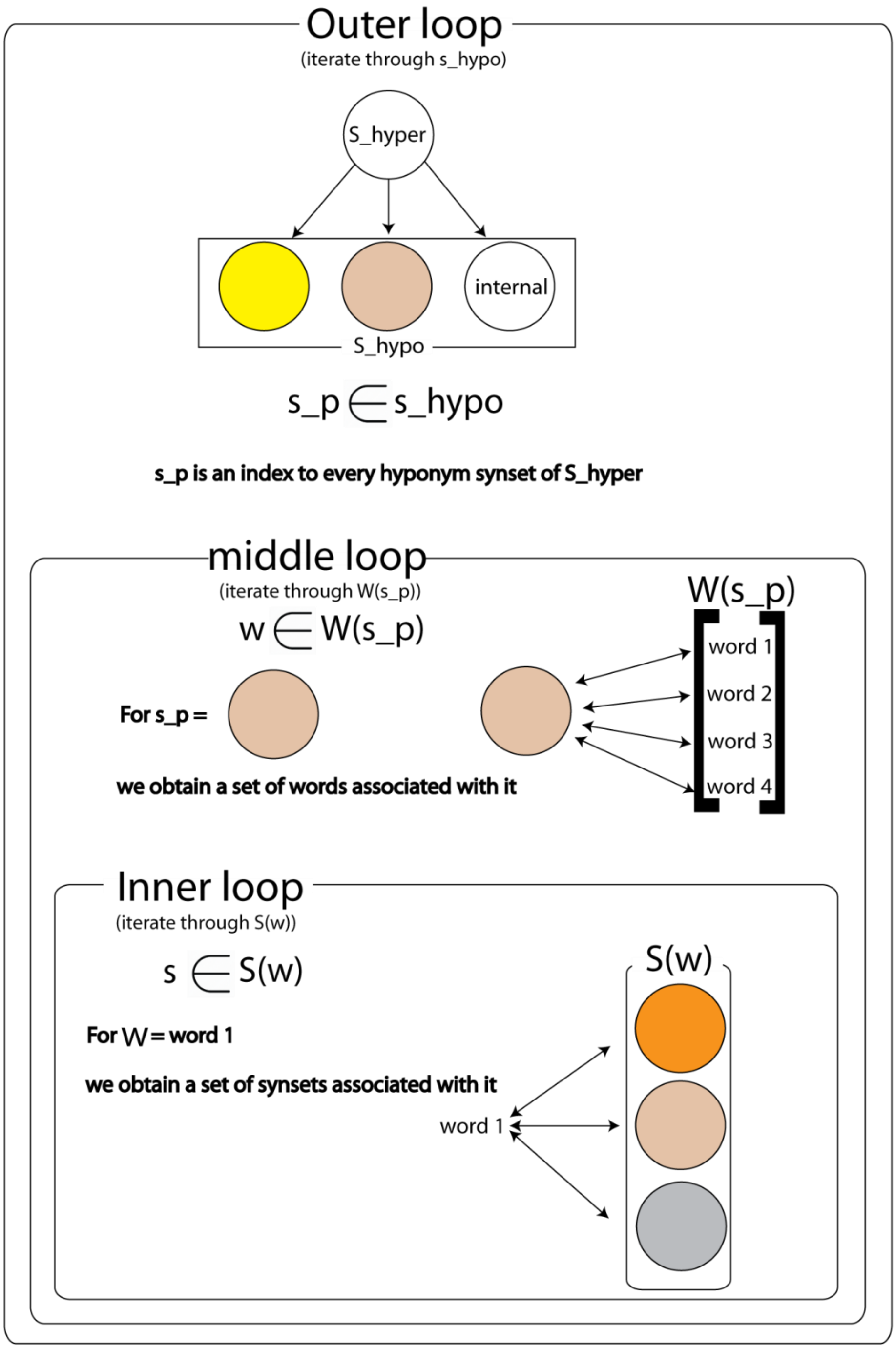
To begin with, it is important to know which synsets are associated with each candidate word. Given that a word can have multiple meanings, it is often associated with more than one synset (many-to-many mapping). WordNet provides a list of synsets associated with a given word (and vice versa). Furthermore, each synset has a frequency value reflecting how often it occurs in a corpus. Instead of directly projecting the VALEX constraint at the lexical level to these synsets, I used this information to weight the VALEX probability of each word by the probability of each associated synset in the list. See Figure A5-2 for an example. This process renders the constraint to be modulated by the actual frequency of the candidate semantic concepts.

But, what if the constraining word ("*climb*") actually prefers a less frequent synset of a lexical item ("*bank*" as "*land alongside a river*")? One might think that this frequency weighting leads to an erroneous projection such that "*bank*" as "*financial institution*" is always more preferred regardless of the context; even in a phrase like "*climb the bank*". In practice, however, many other strongly preferred lexical items generally have the lowest common subsumer at a relatively lower level of the hierarchy with the context-relevant synset regardless of its frequency. The lowest common subsumer refers to the lowest possible hypernym (upper-level synset) that contains all of the input synsets in its hyponym (e.g. "*geological formation*" is the lowest common subsumer of "*river-bank*" and "*mountain*"). Consequently, when it comes to the higher representation at an upper-level of the hierarchy, the semantic preferences are determined mostly by how many candidate synsets fall under the common subsumer. For example, in a context "Hammering the …", a lexical item "*nail*" will be highly preferred which could either mean "*a plate covering a finger*" or "*a metal spike*" with equal frequencies. So, the actual projection will split the lexical preference in half to each of these synsets. However, since many other lexical items preferred by the context will fall under the same category as the "*metal spike*" synset (e.g. "*tool*"), the higher representation at the upper-level (e.g. "*body-part*" vs. "*tool*") will strongly prefer the common subsumer ("tool") even if the projection from "*nail*" equally preferred both "*metal spike*" and "*finger plate*" (see Figure A5-2).

To reflect that all hyponym synsets are embedded within their common subsumer, I accumulated the weighted synset preferences via summation. This straightforward operation ensures that the total preference f at some hypernym synset s_{hyper} takes the preference of itself (represented as an internal node) as well as the preference of each of its hyponym synset (I will denote these synsets as s_{hypo}):

$$f(s_{hyper}) = \sum_{s_p \in s_{\square ypo}} \sum_{w \in W(s_p)} \frac{f(s_p)}{\sum_{s \in S(w)} f(s)} f(w) \ldots (A5.1)$$

where s_p represents every synset contained in s_{hypo}, w represents every lexical item (word) associated with the synset s_p and s represents every synset associated with the word w. W is a function that takes a synset as an input and finds all words associated with it whereas S is an inverse of W such that it takes a word as an input and finds all synsets associated with it. The logic of this equation is visually depicted in Figure A5-1. This procedure of propagating the lexical frequencies into the WordNet hierarchy is depicted in Figure A5-2.



* Notice that $s\_p \in S(w)$ such that

$$\frac{f(s\_p)}{\sum_{s \in S(word1)} f(s)} = \text{relative frequency of s\_p associated with word 1}$$

* Then, $\displaystyle\sum_{w \in W(s\_p)} \frac{f(S\_p)}{\sum_{s \in S(w)} f(s)} f(w) =$ projected lexical frequency at synset S_p

Figure A5-1: A visual description of how to obtain the projected lexical frequency at each synset of interest (see Equation A5.1). See also, Figure A5-2 for propagation of this projected frequency through the WordNet hierarchy.

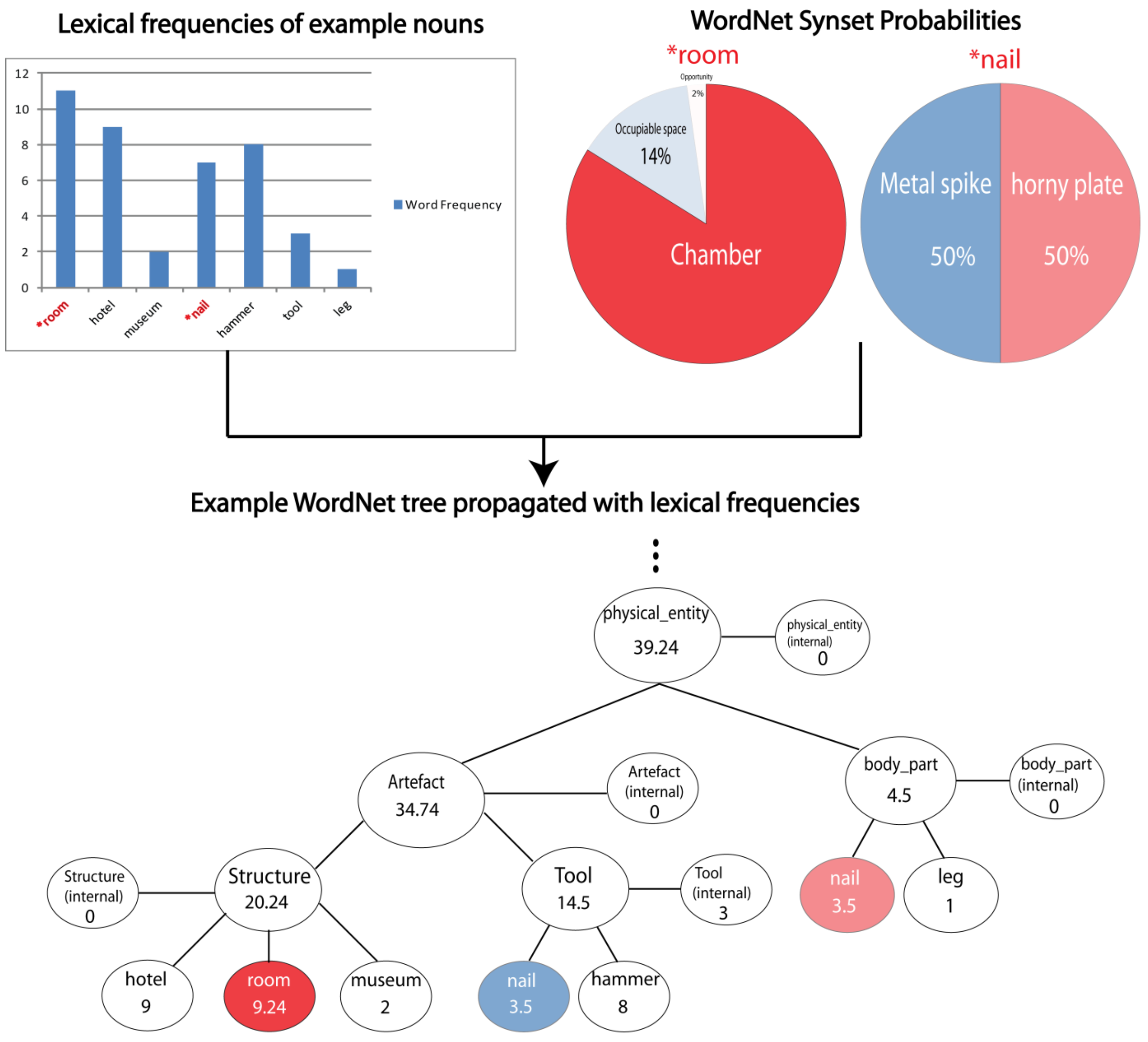


Figure A5-2: A schematic illustration of the propagation of lexical frequencies into the WordNet hierarchy (The actual values are made-up just for the illustration of this process). Note that if a word has multiple synsets associated with it, its lexical frequency is weighted by the relative synset frequency of each of them (e.g. the frequency of "nail" 7 is divided into frequencies of two associated synsets: "metal spike" 3.5 and "finger plate" 3.5). In order to account for the pure frequency of the hypernym sense, the internal node of the hypernym synset was added to the hyponym level (see McCarthy, 2001). This ensured that the

accumulated frequency at the hypernym synset is always equal to the sum of frequencies across all hyponym synsets including its internal synset (see (A5.1)).

However, the fact that the WordNet hierarchy is not a proper tree raises a problem that synsets at the leaves can have multiple paths to the root. This means that the frequency at these hyponym synsets should be shared across multiple paths during the propagation. To address this problem, I applied the same logic to deal with the many-to-many mappings between lexical items and synsets: calculate the probabilistic weight for each path using the frequency information about every synset in the path and apply this weight when the propagation enters this path. The probabilistic weight for each path was computed as follows:

$$p(h_k) = \frac{\sum_{s \in h_k} f(s)}{\sum_{j \in h} \sum_{s \in h_j} f(s)} \ldots (A5.2)$$

where $p(h_k)$ is a probabilistic weight for k^{th} path over a set of paths h that contains synset s. Combining (A5.1) and (A5.2), we have:

$$f(S_{hyper}) = \sum_{s_p \in S_{hyper}} \sum_{w \in s_p} \frac{f(s_p)}{\sum_{s_w \in w} f(s_w)} \frac{\sum_{s_h \in h} f(s_h)}{\sum_{h \in h(s_p)} \sum_{s_h \in h} f(s_h)} f(w) \ldots (A5.3)$$

where s_h is an index to every synset in the path h that the synset s_p belongs to. Note that if there is only one path that s_p belongs to and it is the only one synset associated with a word w, the frequency of this synset in its internal node is simply the lexical frequency $f(w)$. Hence, the original lexical frequency $f(w)$ is essentially modified by the synset probability as well as the path probability that the synset belongs to.

Optimizing the representation: the minimum description length (MDL) principle

Once the lexical frequencies are fully propagated, the entire WordNet hierarchy represents the conceptual space, tuned specifically to the VALEX constraint imposed by the preceding context through (A5.3). However, this space is often inefficient to represent especially because there are many zeros in the leaves (i.e. synsets that are too specific). Further, the representational cost could substantially rise due to many dimensions (synsets) representing the redundant information because a hypernym synset is merely a sum of its hyponym synsets such that it becomes redundant once its hyponym synsets are fully represented. Therefore, modelling the semantic constraint using the entire WordNet hierarchy suffers the problem of representation (see Li & Abe, 1998). To address this problem, I applied the MDL algorithm to find the optimal cut in which the constraint can be efficiently represented in this WordNet conceptual space.

MDL was originally proposed by Jorma Risannen (1978) as a principle of data compression and statistical estimation. It typically consists of two terms (data and parameter description lengths) and finds out the best compromise between them that can minimize their sum. The data description length is a penalty term for compression to prevent the algorithm from compressing data too much and losing variability in the original data space. It is a maximum likelihood estimate (MLE) of a set of parameters θ that maximizes the likelihood of given data S:

$$\hat{\theta} = \max \prod_{x_i \in S} P(x_i|\theta) = \min \sum_{x_i \in S} - log\, P(x_i|\theta) \ldots (A5.4)$$

Of course, the likelihood of S is maximized when there are equal number of parameters as data points (in which case, the model can explain 100% of variance in the data) such that $\theta \in R^N$. This logically renders the algorithm to stay in R^N and, consequently, prevent the data from being compressed. In contrast, the parameter description length penalizes the model for using too many parameters. In other words, it prevents the model from overfitting the data which essentially promotes the compression:

$$\min \frac{k}{2} \log|S| \ldots (A5.5)$$

where $k = |\theta|$ and $|S|$ represents the total number of components in the data. Note that $\frac{1}{2}\log|S|$ is a weight on the number of parameters in the model k: the larger the sample size of the data $|S|$, the more the algorithm favours compression. Therefore, the algorithm initially prefers compression at the leaves of the WordNet hierarchy but such preference diminishes as it gets closer to the root of the hierarchy. This specific form of the weight $\frac{1}{2}\log|S|$ is derived from the fact that the standard deviation of any MLE parameter is approximately $\frac{1}{\sqrt{|S|}}$. As a result, describing each of them using more than $-\log\frac{1}{\sqrt{|S|}} = \frac{1}{2}\log|S|$ bits tends to be wasteful

(Li & Abe, 1998). Recall that a bit (negative log of a probability) is an information unit (see Section 2.4 in Chapter 2).

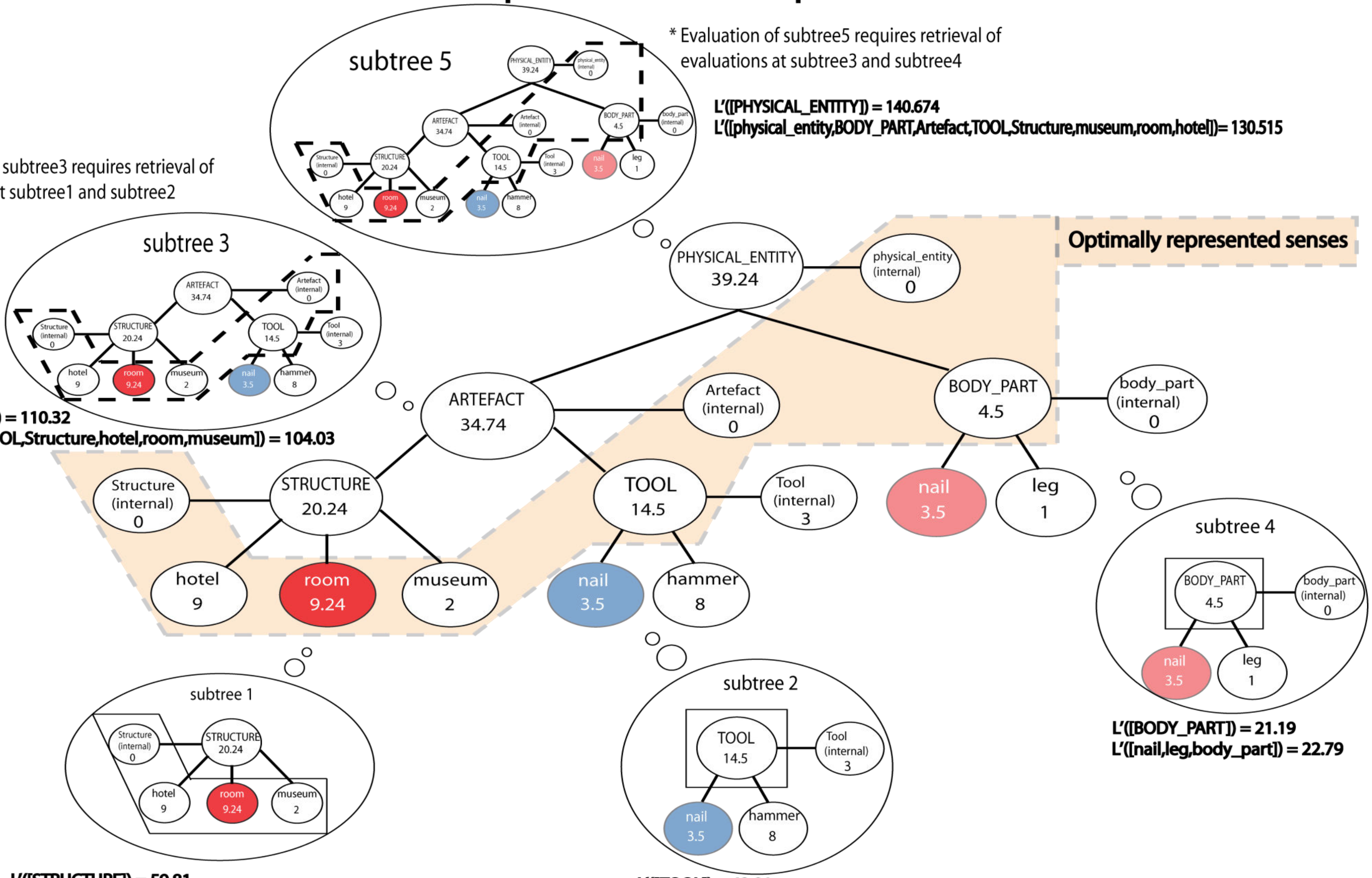
Combining these, the MDL principle is defined as:

$$MDL \coloneqq \min\left(\frac{k}{2}\log|S| + \sum_{i=1}^{N} - \log P(x_i|\theta)\right) \ldots (A5.6)$$

It is worth noting that it is commonly adopted in statistical modelling for the model selection problem such as in a multiple regression to find out the optimal number of predictors to explain the response variable or in auto-regression to choose the model order. In fact, it nicely converges to a very similar solution to the information criterions and its asymptotic behaviour is identical to the Bayesian information criterion (BIC).

However, optimizing the representation of semantic constraint in WordNet is not an easy task because infinite number of models can be generated from the large semantic space in WordNet consisting of 117,000 hierarchically organized synsets. Here, I implemented Li & Abe's subtree evaluation approach in which MDL was compared between the models at hypernym and hyponym levels at every subspace defined by a two-level tree from the leaves. The results of this comparison was saved and later retrieved to evaluate the upper-level trees as it goes down towards the root. The detailed illustration of this procedure is described in Figure A5-3.

MDL-based Optimal semantic representation



* Evaluation of subtree5 requires retrieval of evaluations at subtree3 and subtree4

L'([PHYSICAL_ENTITY]) = 140.674
L'([physical_entity,BODY_PART,Artefact,TOOL,Structure,museum,room,hotel])= 130.515

* Evaluation of subtree3 requires retrieval of evaluations at subtree1 and subtree2

L'([ARTEFACT]) = 110.32
L'([Artefact,TOOL,Structure,hotel,room,museum]) = 104.03

L'([BODY_PART]) = 21.19
L'([nail,leg,body_part]) = 22.79

L'([STRUCTURE]) = 59.81
L'([Structure,hotel,room,museum]) = 54.93

L'([TOOL]) = 43.81
L'([nail,hammer,Tool]) = 46.98

Figure A5-3: A simplistic illustration of how the generalized tree-cut was obtained after the propagation of lexical frequency (see Figure A5-2). For each subtree, I computed the description length L' at both hyponym and hypernym levels and pointed out the level with smaller L'. The level with smaller L' was later retrieved when comparing with the upper hypernym and this procedure was repeated until the algorithm reaches at the root.

In Figure A5-3, the actual description length L' values at the subtree 1 were computed as follows:

$$L^{'}([STRUCTURE]) = -20.24\log\left(\frac{20.24}{39.24} * \frac{1}{4}\right) = 59.8115 \quad \ldots(A5.7)$$

$$\begin{aligned} &L^{'}([structure, hotel, room, museum]) \\ &\quad = -9\log\left(\frac{9}{39.24}\right) - 9.24\log\left(\frac{9.24}{39.24}\right) - 2\log\left(\frac{2}{39.24}\right) + \frac{4-1}{2}\log(39.24) \\ &\quad = 46.9854 + 7.9414 = 54.9268 \quad \ldots(A5.8) \end{aligned}$$

The data (A5.4) and parameter description lengths (A5.5) are marked by purple and green respectively. Recall that $\sum_n P(x) = n * P(x)$ if $P(x)$ is constant across all n (which is the case above in (A5.7) and (A5.8) as the probability of a synset is constant across multiple occurrences of itself). In this pipeline each data sample corresponds to each frequency count (or occurrence of an item) and, as a result, the summation in the data description length is defined over every occurrence of synsets at the hypernym or hyponym level in the subtree. Also, note that the hypernym synset probability $\frac{20.24}{39.24}$ is normalized by the total number of hyponym synsets that it represents (Li & Abe, 1998). This is to ensure that the hypernym synset represents all of its hyponym synsets with equal strengths such that the number of bits to encode the data is represented in a maximally uninformative manner at the hypernym level. As an exchange, the model at the hypernym level is comparably cheaper that the one at the hyponym level because there is only one hypernym at a subtree (i.e. only one parameter in the model) and the hypernym L' is fully determined by the data description length. Note that the initial evaluation automatically elevated to the level where a hypernym synset in a subtree has non-zero frequency count because many specific synsets at the leaves have zero frequency in a large space of 117,000 hierarchically organized synsets.

As another demonstration, the L' values at the subtree 3 was computed as:

$$L'([ARTEFACT]) = -34.74 \log\left(\frac{34.74}{39.24} * \frac{1}{8}\right) = 110.3248 \quad \ldots(31)$$

$$\begin{aligned} L'([artefact, TOOL, structure, hotel, room, museum]) \\ = -14.5 \log\left(\frac{14.5}{39.24} * \frac{1}{3}\right) - 9 \log\left(\frac{9}{39.24}\right) - 9.24 \log\left(\frac{9.24}{39.24}\right) \\ - 2 \log\left(\frac{2}{39.24}\right) + \frac{6-1}{2} \log(39.24) = 90.7933 + 13.2356 \\ = 104.0289 \quad \ldots(32) \end{aligned}$$

Extracting the semantic constraint from the optimal tree-cut Γ

Once the optimal level of representation Γ (with minimum description length) is confirmed (e.g. the region highlighted by orange in Figure A5-3), the last step is to extract a probability distribution from this optimal level. Each synset in this level represents an accumulated frequency value in the hierarchy (see Figure A5-2). Therefore, a simple normalization across these synsets $s_i \in \Gamma \; \forall \; i = 1{:}\,|\Gamma|$ would necessarily render the synset located at the comparably upper-level of the hierarchy to have a higher probability value than the others merely due to its location. This makes the similarity patterns be generally biased and influenced by the synset location in the hierarchy. In order to correct for this accumulative bias, I weighted each synset $s_i \in \Gamma$ by its associated information content (IC) provided by WordNet based on the brown corpus. IC is a negative log of a synset probability (Resnik, 1995), encoding the informativeness (or degree of specificity) of a synset in the hierarchy. In this way, the accumulative bias in Γ can be objectively corrected, assuming that more abstracted synsets contain less information than more specific ones. See Figure A5-4 for extracting the probability of a word represented in Γ. The output IC-weighted probability distribution at Γ was used as a model of the semantic constraint in my analysis.

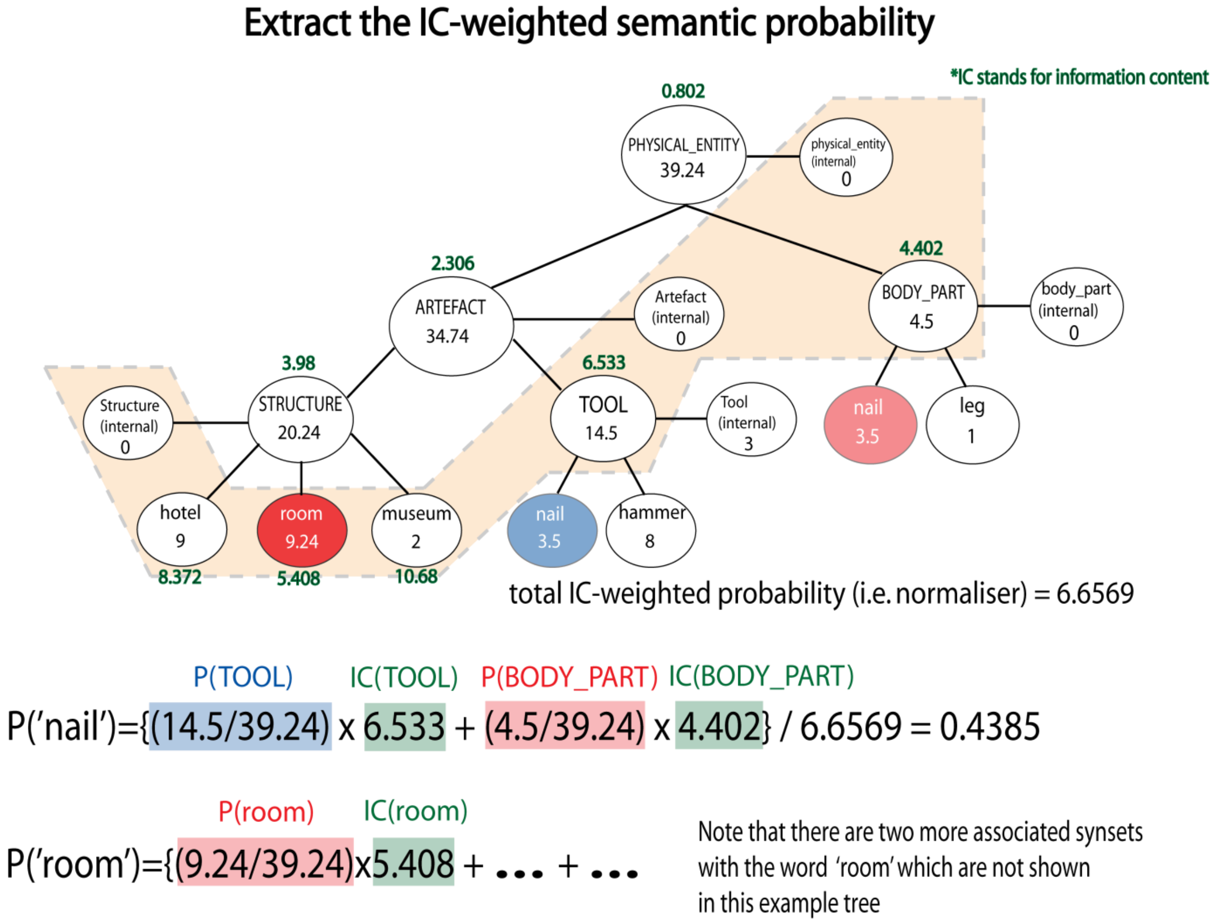

Extract the IC-weighted semantic probability
*IC stands for information content
0.802
PHYSICAL_ENTITY 39.24
physical_entity (internal) 0
2.306
ARTEFACT 34.74
Artefact (internal) 0
4.402
BODY_PART 4.5
body_part (internal) 0
3.98
STRUCTURE 20.24
Structure (internal) 0
6.533
TOOL 14.5
Tool (internal) 3
nail 3.5
leg 1
hotel 9
room 9.24
museum 2
nail 3.5
hammer 8
8.372
5.408
10.68
total IC-weighted probability (i.e. normaliser) = 6.6569
P(TOOL)
IC(TOOL)
P(BODY_PART)
IC(BODY_PART)
P('nail')={(14.5/39.24) x 6.533 + (4.5/39.24) x 4.402} / 6.6569 = 0.4385
P(room)
IC(room)
P('room')={(9.24/39.24)x5.408 + ... + ...
Note that there are two more associated synsets with the word 'room' which are not shown in this example tree

Figure A5-4: A schematic illustration of extracting the IC-weighted semantic probability of a word at the optimal cut. The IC weights were used as objective normalizers for the accumulated probability associated with every synsets contained in the optimal cut. The total IC-weighted probability was calculated as a sum of IC-weighted probability across all synsets in Γ.

Find the mean optimal cut across different optimization schemes

Constraints naturally vary depending on the preceding context. Therefore, the lexical constraints provided by different verbs are always different from each other which lead to different optimal cuts in the WordNet space. This is problematic for RSA analysis which requires trial-wise comparisons based on the information defined in the same space (i.e. the representational geometry must be comparable). The easiest way to address this issue is to concatenate the labelled dimensions across different verbs and remove the repetitions but this approach is not appropriate because it could leave unique senses which are related by hyponymy due to its hierarchical nature of representation. Instead, I proposed a method of finding a mean of the optimized cuts for different verbs through recursive evaluations.

It involves the recursive bottom-up subtree evaluation scheme as described in Figure A5-3 for finding the optimal cut for a given tree. But, instead of using the MDL algorithm with projected lexical constraints, I used the frequency counts of every synset being optimal across 50 different trees associated with each verb in the stimuli (note that the goal of this step is not about finding the parsimonious representation but about finding the optimal cut consisting of the most commonly optimal synsets across different tree). Therefore, I simply counted how many times the hypernym synset is optimal and compared it to the average count of its hyponym synsets being optimal. If the average count was higher, the hyponym synsets were saved and later recalled when evaluating at a subtree in the upper hierarchy. Through this scheme, the mean optimal cut was found, consisting of 15 different synsets. As an output illustration, Figure 2-8 in Chapter 2 shows the semantic constraints of different verbs defined by these 15 synsets.

A5-references

Li, H., & Abe, N. (1998). Generalizing case frames using a thesaurus and the MDL principle. *Computational Linguistics*, *24*(2), 217-244.

McCarthy, D. (2001). *Lexical acquisition at the syntax-semantics interface: diathesis alternations, subcategorization frames and selectional preferences* (Doctoral dissertation, University of Sussex).

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*.

Risannen, J. (1978). Modelling by shortest data description. *Automatica*, *14*(5), 465-471.

Appendix 6: Derivatives of the non-linear functions

One of the common aspects of the non-linear functions introduced in the main text is that they all have a nice and simple derivative. This renders the gradient computation easier, often leading to faster learning.

6-1. Derivative of sigmoid

Provided $Z = sigmoid(q)$, how do we compute $\frac{dZ}{dq}$?

$$Z = \frac{1}{1+e^{-q}} = (1+e^{-q})^{-1}$$

Hence,

$$\frac{dZ}{dq} = (-1)(1+e^{-q})^{-2}\frac{d}{dq}(1+e^{-q}) = -\frac{1}{(1+e^{-q})^2}(-e^{-q}) = \frac{e^{-q}}{(1+e^{-q})^2}$$

Here, we can introduce a trick:

$$\frac{dZ}{dq} = \frac{(1+e^{-q}-1)}{(1+e^{-q})^2} = \frac{1+e^{-q}}{(1+e^{-q})^2} - \frac{1}{(1+e^{-q})^2} = \frac{1}{1+e^{-q}} - \frac{1}{1+e^{-q}}\frac{1}{1+e^{-q}}$$

Factor out Z and obtain:

$$\frac{dZ}{dq} = \frac{1}{1+e^{-q}}\left(1-\frac{1}{1+e^{-q}}\right) = Z(1-Z)$$

6-2. Derivative of ReLU

Provided $Z2 = ReLU(q) = \max(0,q) = \begin{cases} q \; if \; q > 0 \\ 0 \; if \; q \leq 0 \end{cases}$, the derivative is straightforward to compute:

$$\frac{d}{dq}Z2 = \begin{cases} 1 \; if \; q > 0 \\ 0 \; if \; q \leq 0 \end{cases}$$

6-3. Derivative of hyperbolic tangent

One of the trigonometric property of the hyperbolic tangent allows its derivative to be expressed as following:

$$\frac{d}{dq}Z3 = \frac{d}{dq}\frac{\sinh(q)}{\cosh(q)}$$

Let $f(q) = \sinh(q)$ and $g(q) = \cosh(q)$. Using the quotient rule, $\left[\frac{f(q)}{g(q)}\right]' = \frac{g(q)f'(q) - f(q)g'(q)}{[g(q)]^2}$, we can express the derivative as:

$$\frac{d}{dq}Z3 = \frac{\cosh(q)\frac{d}{dq}\sinh(q) - \sinh(q)\frac{d}{dq}\cosh(q)}{\cosh^2(\mathrm{q})}$$

where $\frac{d}{dq}\sinh(q) = \cosh(q)$ and $\frac{d}{dq}\cosh(q) = \sinh(q)$. Hence, the expression simplifies to:

$$\frac{d}{dq}Z3 = \frac{\cosh^2(\mathrm{q}) - \sinh^2(\mathrm{q})}{\cosh^2(\mathrm{q})} = 1 - \tanh^2(\mathrm{q})$$

6-4. Derivative of softplus

Provided $Z4 = softplus(q) = \log(1 + e^q)$, its derivative is easily visible as:

$$\frac{d}{dq}Z4 = \frac{1}{1 + e^q} * e^q = \frac{1}{1 + e^{-q}}$$

Highlight that this softplus derivative is exactly same as the sigmoid function described above.

6-5. Derivative of softmax

Let $f(a) = e^{a_j}$ and $g(a) = \sum_k e^{a_k}$ where $o_j = \frac{e^{a_j}}{\sum_k e^{a_k}}$. Then, softmax function could be expressed as a ratio between these two functions as $\frac{f(a)}{g(a)} = \frac{e^{a_j}}{\sum_k e^{a_k}}$. According to quotient rule, $\left[\frac{f(a)}{g(a)}\right]' = \frac{g(a)f'(a) - f(a)g'(a)}{[g(a)]^2}$. Hence, softmax derivative could be expressed as:

$$\frac{\partial o_j}{\partial a_i} = \frac{\sum_k e^{a_k}\left[\frac{\partial}{\partial a_i}(e^{a_j})\right] - e^{a_j}\left[\frac{\partial}{\partial a_i}(\sum_k e^{a_k})\right]}{[\sum_k e^{a_k}]^2}$$

Each of the derivative terms is expressed as:

$$\frac{\partial}{\partial a_i}(e^{a_j}) = I_{i=j}e^{a_j}$$

$$\frac{\partial}{\partial a_i}\left(\sum_k e^{a_k}\right) = e^{a_{k=i}} = e^{a_i}$$

where $I_{i=j}$ is an indicator function which assigns 1 if $i = j$ or 0 otherwise. Substituting these provides:

$$\frac{\partial o_j}{\partial a_i} = \frac{\sum_k e^{a_k}\left[I_{i=j} e^{a_j}\right]}{\left[\sum_k e^{a_k}\right]^2} - \frac{e^{a_j}\left[e^{a_i}\right]}{\left[\sum_k e^{a_k}\right]^2} = \frac{I_{i=j} e^{a_j}}{\sum_k e^{a_k}} - \frac{e^{a_j}}{\sum_k e^{a_k}} \frac{e^{a_i}}{\sum_k e^{a_k}}$$

Provided $o_j = \frac{e^{a_j}}{\sum_k e^{a_k}}$ and $o_i = \frac{e^{a_i}}{\sum_k e^{a_k}}$:

$$\frac{\partial o_j}{\partial a_j} = I_{i=j} o_j - o_j o_i = \begin{cases} o_{j=i}(1 - o_i) & if\ i = j \\ -o_j o_i & if\ i \neq j \end{cases}$$

Appendix 7: Backpropagation and gradient learning

From 2.3.1 and 2.3.2 in Chapter 2, the functional architecture of the system is constructed. Now, we just need to train this system through the data we prepared so that it can learn the statistical (non-linear) patterns to generate as an accurate response as possible. Following on from Figure 2-3, the system generates an output O which can be evaluated against the data. Therefore, the first step of designing a training algorithm for this system is to define a loss function. Throughout this thesis, I set the sigmoid and softmax as default hidden and output layer activation functions respectively because these functions are used in the neural network that I use for language modelling.

The softmax output function is paired with the cross entropy loss. This is because maximizing the log likelihood of the softmax classification (or multinomial logistic regression) is same as minimizing the cross entropy of the actual and predicted distributions. In the context of training a neural network, the loss can be expressed as:

$$H(Y,O) = -\sum_{j=1}^{J} y_j \ln y_j + \sum_{j=1}^{J} y_j \ln \frac{y_j}{o_j} = \sum_{j=1}^{J} Y_j \left(\ln \frac{y_j}{o_j} - \ln y_j \right) = -\sum_{j=1}^{J} y_j \ln o_j \ldots (A7.1)$$

where J is the total number of output classes (or neurons). Given the binary response vector Y whose entropy is zero (i.e. one-hot vector), the objective simplifies to minimizing the difference between the actual response and the output of the network quantified by the Kullback-Leibler divergence. Note that I assume O as an output vector for simpler illustration with an assumption of N = 1 in Figure 2-3.

As in the other typical classification algorithms, the optimization problem is to find weights $W1$ and $W2$ that minimize the loss function (A7.1). However, the input underwent a number of transformations through different neurons in different layers to generate the output. Therefore, the optimization involves back-propagating the error from the output to the input layer so that the network can adjust the weights accordingly. This can be formulated using the chain rule as below:

$$\frac{\partial}{\partial s1_j} H(Y,O) = \sum_{k=1}^{J} \frac{\partial H(Y,O)}{\partial o_k} \frac{\partial o_k}{\partial s1_j} \ldots (A7.2)$$

$$\frac{\partial}{\partial W2_{qj}} H(Y,O) = \frac{\partial H(Y,O)}{\partial s1_j} \frac{\partial s1_j}{\partial W2_{qj}} \ldots (A7.3)$$

$$\frac{\partial}{\partial W1_{pq}}H(Y,O)=\sum_{j=1}^{J}\frac{\partial H(Y,O)}{\partial s1_j}\frac{\partial s1_j}{\partial z_q}\frac{\partial z_q}{\partial s2_q}\frac{\partial s2_q}{\partial W1_{pq}}\ldots(A7.4)$$

where p, q and j are indices of the neurons in the input, hidden and output layers respectively, $s1$ is an input to the output layer defined as $Z*W2+b2$ in Figure 2-3 and $s2$ is an input to the hidden layer defined as $X*W1+b1$ in Figure 2-3. The summation across multiclass J in (A7.2) reflects that the output O is normalized by activation values of the other neurons in the layer; hence, the activation at jth output neuron does not solely depend on its input. This is why the error gradient at the other output neurons must be integrated to compute the gradient at the input to the jth output neuron. Each term in (A7.3) can be computed as follows:

$$\frac{\partial H(Y,O)}{\partial o_j}=\frac{\partial}{\partial o_j}\left(-\sum_{j=1}^{J}y_j\log o_j\right)=-\frac{y_j}{o_j}\ldots(A7.5)$$

$$\begin{aligned}\frac{\partial H(Y,O)}{\partial s1_j}&=\sum_{k=1}^{J}\frac{\partial H(Y,O)}{\partial o_k}\frac{\partial o_k}{\partial s1_j}=\frac{\partial H(Y,O)}{\partial o_j}\frac{\partial o_j}{\partial s1_j}+\sum_{k\neq j}\frac{\partial H(Y,O)}{\partial o_k}\frac{\partial o_k}{\partial s1_j}\\&=-\frac{y_j}{o_j}\left(o_j(1-o_j)\right)-\sum_{k\neq j}\frac{y_k}{o_k}(-o_jo_k)=-y_j(1-o_j)+o_j\sum_{k\neq j}y_k\\&=-y_j+o_j\left(y_j+\sum_{k\neq j}y_k\right)=-y_j+o_j\left(\sum_{k}y_k\right)=o_j-y_j\ldots(A7.6)\end{aligned}$$

$$\frac{\partial s1_j}{\partial W2_{qj}}=\frac{\partial}{\partial W2_{qj}}\left(\sum_{q=1}^{Q}z_qW2_{qj}+b2_j\right)=z_q\ldots(A7.7)$$

where Q is the total number of neurons in the hidden layer. See Appendix 6 for a proof of the softmax derivative. Hence, putting (A7.6) and (A7.7) together provides:

$$\frac{\partial}{\partial W2_{qj}}H(Y,O)=z_q\left(o_j-y_j\right)\ldots(A7.8)$$

Similarly, each weight in $W1$ can be updated as follows:

$$\frac{\partial s1_j}{\partial z_q} = \frac{\partial}{\partial z_q}\left(\sum_{q=1}^{Q} z_q W2_{qj} + b2_j\right) = W2_{qj} \dots (A7.9)$$

$$\frac{\partial z_q}{\partial s2_q} = \frac{\partial}{\partial s2_q}\left(\frac{1}{1+e^{-s2_q}}\right) = z_q(1-z_q) \dots (A7.10)$$

$$\frac{\partial s2_q}{\partial W1_{pq}} = \frac{\partial}{\partial W1_{pq}}\left(\sum_{p=1}^{P} x_p W1_{pq} + b1_q\right) = x_p \dots (A7.11)$$

where P is the total number of neurons in the input layer. Combining (A7.6), (A7.9), (A7.10) and (A7.11) provides:

$$\frac{\partial}{\partial W1_{pq}} H(Y,O) = \sum_{j=1}^{J} W2_{qj}(o_j - y_j) z_q (1-z_q) x_p \dots (A7.12)$$

The equations (A7.8) and (A7.12) show how connectivity patterns in a network are modified as a function of experience. Referring to the parallel distributed processing (see section 4.2 in Chapter 4) framework where $\Delta w_{i,j} = g\big(a_i(t), t_i(t)\big)h\big(o_j(t), w_{i,j}\big)$, the updating expression of (A7.8) can be expressed in this form by specifying the arbitrary functions g and h such that $g(o_j, y_j) = (o_j - y_j)$ and $h(z_q, W2_{qj}) = z_q$. The updating expression of (A7.12) can be expressed in an expanded form by specifying more arbitrary functions l and f as following: $g(o_j, y_j) = (o_j - y_j)$, $h(z_q, W2_{qj}) = W2_{qj}$, then, $l(s2_q, z_q) = z_q(1-z_q)$ and $f(x_p, W1_{pq}) = x_p$. Additional functions are necessary because the gradient (teaching materials) passes through the hidden layer. Similar to y_j working as a teacher in g, z_q works as a teacher in l modifying the output from the input layer $s2_q$. Therefore, all these implementations of weight updating fit well with the traditional Hebbian learning, strengthening the connectivity between neurons which are firing together to generate an accurate response.

So far, I described how the error gradient can be propagated back to the different layers (I used the sigmoid and softmax as activation functions in the hidden and output layers respectively as an example but the same logic can be applied with different activation functions). Now, I briefly describe one of the most popular optimization algorithms, gradient descent, to implement the modification of the connectivity patterns $W1$ and $W2$. Gradient descent is often used to minimize a loss function (in this case, the cross entropy parameterized by these patterns) by updating the parameters in the opposite direction of the gradient of the loss function:

$$W2_{qj} \coloneqq W2_{qj} - \eta * \frac{\partial}{\partial W2_{qj}} H(Y, O) \ldots (A7.13)$$

$$W1_{pq} \coloneqq W1_{pq} - \eta * \frac{\partial}{\partial W1_{pq}} H(Y, O) \ldots (A7.14)$$

where η is a learning parameter that determines the speed-accuracy tradeoff in finding a (local) minimum. For example, with high η, the algorithm can rapidly search for the minimum by taking a large step towards the minimum but, if the step is too large, it might overlook the minimum and find itself difficult to converge. Therefore, it is important to set η properly for efficient optimization.

There are a number of variants available in practice depending on the amount of data used to compute the gradient before updating. On the one hand, it is possible to use the entire dataset to compute as stable gradient as possible before updating. On the other hand, the patterns can be updated for every sample based on the unstable gradient computed from one sample. Not surprisingly, the first approach (called batch gradient descent) is often infeasible due to the amount of time it takes to converge whereas the second approach (called stochastic gradient descent) often overshoots and jumps out from the (local) minima due to the fluctuating gradient with high variance (although this can be controlled using the learning rate parameter η). In a midway between these two extremes, one can split the dataset into chunks and update the patterns for each chunk. This is known as mini-batch gradient descent which is designed to reach the convergence in more stable manner while being time-efficient. Hence, the updates can be expressed as (from (A7.13) and (A7.14)):

$$W2_{qj} \coloneqq W2_{qj} - \eta \frac{1}{N} \sum_{i=1}^{N} \frac{\partial}{\partial W2_{qj}} H(Y, O) \ldots (A7.15)$$

$$W1_{pq} \coloneqq W1_{pq} - \eta \frac{1}{N} \sum_{i=1}^{N} \frac{\partial}{\partial W1_{pq}} H(Y, O) \ldots (A7.16)$$

where N is the total number of samples in the mini-batch (chunk). It simply computes the average gradient across the entire samples in the mini-batch and updates the connectivity patterns by subtracting the averaged gradient from the connectivity (or weight) matrix (i.e. taking the opposite direction of the gradient of the loss function).