Imaging Neural Correlates of Syntactic Complexity in a Naturalistic Context

by

Asaf Bachrach

Submitted to the Department of Linguistics and Philosophy in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Linguistics at the MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2008

© Asaf Bachrach, MMVIII. All rights reserved.

The author hereby grants to MIT permission to reproduce and distribute publicly paper and electronic copies of this thesis document in whole or in part.

[Signature]

Author

[Signature]

the Department of Linguistics and Philosophy

August 5, 2008

Certified by

[Signature]

John Gabrieli

Professor, Department of Brain and Cognitive Sciences
Thesis Supervisor

Certified by

[Signature]

Alec Marantz

Professor, Department of Linguistics, NYU
Thesis Supervisor

Accepted by

[Signature]

Irene Heim

Head, Department of Linguistics and Philosophy
Abstract

The aim of this thesis, and the research project within which it is embedded, is to delineate a neural model of grammatical competence. For this purpose, we develop here a novel integrated, multi-disciplinary experimental paradigm that endorses the fundamental premise of generative grammar, that the study of language is in essence, the study of the mind.

We use functional Magnetic Resonance Imaging (fMRI) to monitor brain activation while subjects listen to short narratives. The texts have been written so as to introduce various syntactic complexities (relative clauses, embedded questions, etc.) not usually found (in such density) in actual corpora. We have calculated a number of complexity measures (both at the level of the single word and at that of the phrase) based on current linguistic and psycholinguistic theory and with the use of a computationally implemented probabilistic parser. By correlating these measures with observed brain activity, we are able to identify the different brain networks that support linguistic processing and characterize their particular function. Conversely, we use the rich brain data to inform our cognitive, and linguistic, theory.

We report here the neural correlates of surprisal (based on contextual predictions), syntactic complexity, structural ambiguity and disambiguation, Theory of Mind and non-local dependencies. This work made use of novel solutions to compute numerical predictions for these linguistic dimensions, which are often tested only qualitatively, and of a novel parametric fMRI design that allowed for the use of single subject un-averaged data as the dependent variable. The thesis ends with a synthesis of the results in the form of a blueprint for a neural model of grammatical competence.
To my family
Acknowledgments

First and foremost I would like to thank Alec Marantz, which has been not only my supervisor and role model, but the reason for my decision to come to MIT, and the person who introduced me to the cognitive neuroscience of language. I am dearly indebted to John Gabrieli, who has been kind enough to host me in his lab for the last two years and whose generosity made this project possible. As a co-chair of my committee, his foresighted suggestions and continuous feedback and support have been priceless. The many hours of discussion with David Pesetsky, Danny Fox and Matti Hamalainen, the remaining members of my committee, have had a significant impact on the shape of this thesis and in particular the framing of the theoretical issues implicated by this work.

I cannot imagine how this thesis would have come out (if at all) without Brian Roark's collaboration. Working with Brian has been a true pleasure and a continuous source of learning. Brian's enthusiasm for the project and generosity allowed me to explore aspects of computational linguistics I would not have otherwise, had any access to.

Very special thanks go to Carlos Cardenas, my undergraduate research assistant who has become, through his committed work and motivated spirit, a full collaborator on the experiment reported here and other related on-going projects. Two other major contributors to this experiment have been Kate Stoel Gammon, who volunteered to write the narratives used here and Jonah Katz, who did a significant amount of editing and linguistic analysis and in addition contributed his voice.

A great number of people helped me to navigate the complex ocean of fMRI. Special thanks go to Susan Whitfield-Gabrieli, who guided me through all the stages of this project and Eric Zarahn who volunteered many of his hours towards the ungrateful task of teaching me linear algebra and fMRI statistics. I would also like to thank Yuval Nir, Jon Brennan, Thomas Witzel, Dafna Palti, Itamar Kahn, Satra

---

1I am not sure if there is a page limit on the acknowledgment section, but just in case, I will restrict my comments here to the thesis project and will not attempt to acknowledge all the (many) persons who had a formative effect on my academic path before and during my years at MIT.
Ghosh and Oliver Hinds and Olga Felsovalyi for their technical help and advice. I had the fortune to have the excellent support of the Martinos MIT MRI team.

The work here benefited enormously from conversations with some very smart people from a variety of fields. A partial list includes Michael Collins, David Caplan, Ted Gibson, John Hale, Shravan Vasishth, Tim O’Donnell, Roni Katzir and Raj Singh.

Finally, a special thanks to Vito and *La Cachaca* for not giving up on me.
Contents

1 Introduction ................................................. 19
   1.1 ‘A mirror of mind’ (Chomsky, 1975b) .................. 19
       1.1.1 The triple C: Competence, Cognition and Computation 20
   1.2 The use of fMRI in the study of language and sentence processing 21
       1.2.1 The subtractive paradigm .......................... 22
       1.2.2 Deviationism ...................................... 24
       1.2.3 Meta-linguistic tasks and other issues of (un-)naturality ..... 25
   1.3 A triple C paradigm ..................................... 27
       1.3.1 A parametric design .............................. 27
       1.3.2 Theoretical and computational measures .............. 29
       1.3.3 A naturalistic task and context ........................ 31
   1.4 The structure of the thesis ............................ 32
   1.5 Methods .................................................. 33
       1.5.1 Participants ........................................ 33
       1.5.2 Procedure .......................................... 33
       1.5.3 Materials .......................................... 34
       1.5.4 Data acquisition ................................... 39
       1.5.5 Data preprocessing .................................. 40
       1.5.6 Data analysis ..................................... 40
   1.6 Behavioral results ..................................... 48
   1.7 FMRI results ............................................ 49
       1.7.1 Word rate ......................................... 49
1.7.2 Open class word length ........................................... 51
1.8 Discussion ............................................................. 54
  1.8.1 Word rate ....................................................... 54
  1.8.2 Length ............................................................ 55

2 Expectation and surprisal .............................................. 57
  2.1 Introduction ....................................................... 57
    2.1.1 Pre-activation ............................................... 57
    2.1.2 Familiarity and predictability ............................ 58
    2.1.3 Lexical and syntactic pre-activation ...................... 59
    2.1.4 Surprise and uncertainty ................................... 60
    2.1.5 Outline of the chapter ..................................... 60
  2.2 The source of pre-activation .................................... 61
    2.2.1 Frequency and predictability .............................. 61
    2.2.2 Distinguishing prediction and familiarity ............... 66
  2.3 Syntactic versus lexical prediction ............................ 67
    2.3.1 Surprisal ..................................................... 68
    2.3.2 Syntactic versus lexical surprisal ....................... 69
    2.3.3 Surprisal and syntactic complexity ...................... 73
  2.4 Surprise and entropy ............................................. 74
    2.4.1 Syntactic entropy and complexity ....................... 76
  2.5 Brain correlates of expectation during sentence processing .... 77
    2.5.1 The N400 ..................................................... 77
    2.5.2 N400 and the late positivity .............................. 81
    2.5.3 The Early Left Anterior Negativity (ELAN) ............... 82
    2.5.4 FMRI ......................................................... 83
  2.6 The current experiment ........................................... 86
    2.6.1 Correlation across the different measures ............... 88
  2.7 Results: Word class model ....................................... 89
    2.7.1 Neural correlates of lexical surprisal and frequency .... 89
2.7.2 Neural correlates of syntactic surprisal and number of derivation steps .................................................. 93

2.8 Discussion: Word class model .................................................. 96
2.8.1 The neural correlates of lexical surprise .................................. 96
2.8.2 The neural correlates of syntactic surprisal and derivation steps 101

2.9 Results: The Sentence model .................................................. 105
2.9.1 Neural correlates of lexical entropy and averaged frequency .............................................................. 105
2.9.2 Neural correlates of syntactic entropy and averaged number of derivation steps ......................................... 107

2.10 Discussion: Sentential measures .................................................. 110
2.10.1 Lexical entropy and averaged frequency .................................................. 110
2.10.2 Syntactic entropy and syntactic complexity .................................................. 111

2.11 General discussion .................................................. 112

3 Ambiguity 115

3.1 Structural ambiguity in sentence processing .................................................. 115
3.1.1 Parallel versus serial parsing .................................................. 117
3.1.2 Deterministic versus probabilistic parsing .................................................. 117
3.1.3 Sources of information in ambiguity resolution .................................................. 118
3.1.4 Tolerating indeterminacy .................................................. 119

3.2 Evidence for a probabilistic parallel tolerant parser .................................................. 121
3.2.1 Evidence that the parser computes multiple analyses .................................................. 121
3.2.2 Lack of competition effects during the ambiguous region .................................................. 123
3.2.3 Ambiguity is not always resolved .................................................. 125
3.2.4 Evidence from disambiguation cost .................................................. 130
3.2.5 A note on modality .................................................. 133
3.2.6 Summary .................................................. 133

3.3 The current experiment .................................................. 134
3.3.1 The ambiguity measure: beam entropy .................................................. 134
3.3.2 Disambiguation towards a less preferred parse: Re-ranking .................................................. 135
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.4</td>
<td>Results</td>
<td>137</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Neural correlates of beam entropy</td>
<td>137</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Neural correlates of re-ranking</td>
<td>139</td>
</tr>
<tr>
<td>3.5</td>
<td>Discussion</td>
<td>141</td>
</tr>
<tr>
<td>3.5.1</td>
<td>Ambiguity</td>
<td>141</td>
</tr>
<tr>
<td>3.5.2</td>
<td>Re-ranking</td>
<td>143</td>
</tr>
<tr>
<td>3.6</td>
<td>General discussion</td>
<td>144</td>
</tr>
<tr>
<td>4</td>
<td>Theory of Mind</td>
<td>147</td>
</tr>
<tr>
<td>4.1</td>
<td>The competence performance distinction in the study of TOM</td>
<td>148</td>
</tr>
<tr>
<td>4.2</td>
<td>Intentions vs. beliefs</td>
<td>150</td>
</tr>
<tr>
<td>4.3</td>
<td>TOM and language</td>
<td>154</td>
</tr>
<tr>
<td>4.4</td>
<td>A note on neuro-imaging of TOM</td>
<td>157</td>
</tr>
<tr>
<td>4.5</td>
<td>The current experiment</td>
<td>158</td>
</tr>
<tr>
<td>4.5.1</td>
<td>Identification of the TOM network</td>
<td>159</td>
</tr>
<tr>
<td>4.5.2</td>
<td>Distinguishing intentions and beliefs</td>
<td>160</td>
</tr>
<tr>
<td>4.5.3</td>
<td>The relation between modality and TOM</td>
<td>161</td>
</tr>
<tr>
<td>4.5.4</td>
<td>TOM and syntax</td>
<td>163</td>
</tr>
<tr>
<td>4.6</td>
<td>Results</td>
<td>163</td>
</tr>
<tr>
<td>4.6.1</td>
<td>Sentence level effects of TOM</td>
<td>163</td>
</tr>
<tr>
<td>4.6.2</td>
<td>Brain correlates of intentions and belief states</td>
<td>167</td>
</tr>
<tr>
<td>4.6.3</td>
<td>Brain correlates of modality and TOM</td>
<td>172</td>
</tr>
<tr>
<td>4.7</td>
<td>Discussion</td>
<td>176</td>
</tr>
<tr>
<td>4.7.1</td>
<td>Identification of the Theory of Mind network</td>
<td>176</td>
</tr>
<tr>
<td>4.7.2</td>
<td>Intentions and beliefs</td>
<td>177</td>
</tr>
<tr>
<td>4.7.3</td>
<td>TOM and modality</td>
<td>181</td>
</tr>
<tr>
<td>5</td>
<td>Non-local dependencies</td>
<td>183</td>
</tr>
<tr>
<td>5.1</td>
<td>The neurolinguistics of non-local dependencies</td>
<td>184</td>
</tr>
<tr>
<td>5.1.1</td>
<td>The A/A-bar distinction</td>
<td>184</td>
</tr>
<tr>
<td>5.1.2</td>
<td>Maintenance and retrieval of A-bar dependencies</td>
<td>188</td>
</tr>
</tbody>
</table>
A.0.2 Surprisal ........................................ 239
A.0.3 Lexical and Syntactic surprisal ................... 239
A.0.4 Derivation steps .................................. 240

B The narratives ........................................ 243
B.1 Narrative I ........................................... 243
B.2 Narrative II .......................................... 245
B.3 Narrative III ......................................... 247
B.4 Narrative IV .......................................... 250
List of Figures

1-1 Correlation matrix across the 13 parametric modulators of the word condition in the Word model ........................................ 45
1-2 Correlation matrix across the 9 parametric modulators of the open class condition in the Word-Class model ............................ 46
1-3 Correlation matrix across the 9 parametric modulators of the sentence condition in the Sentence model .............................. 47
1-4 Positive and negative effects of word rate ................................ 52
1-5 Positive effects of word length ........................................... 53

2-1 Effects of lexical surprisal and frequency ................................ 92
2-2 Effects of number of derivation steps and syntactic surprisal .... 95
2-3 Effects of lexical entropy and averaged frequency .................... 106
2-4 Effects of syntactic entropy and complexity ............................ 109

3-1 Reading times from Swets et al. (2008) .................................. 128
3-2 Question response times from Swets et al. (2008) ....................... 128
3-3 Effects of beam entropy ................................................... 138
3-4 Effects of re-ranking ....................................................... 140

4-1 Positive correlation with sentential TOM ................................ 166
4-2 Positive correlation with the two word level TOM measures ........ 171
4-3 Effects of the modal measure and the contrast TOM-modal ........ 175

5-1 Positive correlation with the sentence model displacement parameter 195
5-2 Positive correlation with word-level model displacement measures .. 199
6-1 A graphic summary of the grammatical competence model . . . . . 224
List of Tables

1.1 Word rate results (positive) ........................................... 49
1.2 Word rate results (negative) ........................................... 50
1.3 Word length results ..................................................... 51

2.1 Correlation values (r) between surprisal, frequency and length measures: open class words ........................................ 89
2.2 Lexical surprisal results .................................................. 90
2.3 Log frequency results .................................................... 90
2.4 Combined lexical surprisal and frequency results .................. 91
2.5 Derivation steps results .................................................. 93
2.6 Syntactic surprisal results .............................................. 93
2.7 Derivation steps and syntactic surprisal combined .................. 94
2.8 Sentence lexical entropy results ....................................... 105
2.9 Sentence averaged frequency results .................................. 105
2.10 Combined sentence entropy and averaged frequency .............. 106
2.11 Sentence syntactic entropy results .................................... 107
2.12 Sentence syntactic complexity results ............................... 107
2.13 Sentence combined syntactic entropy and complexity results .... 108

3.1 Beam entropy results ..................................................... 137
3.2 re-ranking results ........................................................ 139

4.1 TOM Results ............................................................ 164
4.2 TOM Results: belief states ............................................. 167
4.3 TOM Results: intentions .................................................. 168
4.4 TOM Results: beliefs - intentions ........................................ 169
4.5 TOM Results: intentions- beliefs ......................................... 169
4.6 Modal Results ............................................................... 172
4.7 TOM-modal Results ......................................................... 173
5.1 Displacement Results ....................................................... 194
5.2 The maintenance measure Results ....................................... 196
5.3 The A-bar gap measure results .......................................... 197
5.4 The A gap measure results ............................................... 198
6.1 Components of the grammatical competence model and their anatom- ical implementation ..................................................... 223
Chapter 1

Introduction

1.1 ‘A mirror of mind’ (Chomsky, 1975b)

In *Reflections on Language* (Chomsky, 1975b), Chomsky, in considering the question *why study language*, writes:

> By studying the properties of natural languages, their structure, organization, and use, we may hope to gain some understanding of the specific characteristics of human intelligence. We may hope to learn something about human nature; something significant, if it is true that human cognitive capacity is the truly distinctive and most remarkable characteristic of the species. Furthermore, it is not unreasonable to suppose that the study of this particular human achievement, the ability to speak and understand a human language, may serve as a suggestive model for inquiry into other domains of human competence and action that are not quite so amenable to direct investigation (p. 4-5).

This view, that the study of language is, fundamentally, a study of the mind, is at the heart of the work reported here. We take this view to entail a strong and direct connection between the theory of grammar, as studied by linguists, and the neural architecture supporting language. In other words, we take the theory of grammar to be a theory about brain organization and, conversely, we take data regarding neural structure and function to bear directly on our theory of grammar.
1.1.1 The triple C: Competence, Cognition and Computation

It appears that in the 1950's and 1960's – the early days of Generative grammar, the (modern) cognitive sciences, and computational linguistics – the study of language was conceived of as an integrated project, shared by all three fields. For a variety of reasons, during the 1970's and 80's, this integrative view lost currency. Instead, each of these three disciplines appropriated to itself a sub-domain of the study of language. Linguistics became the study of grammatical competence, or speakers’ knowledge of language. Psycholinguistics emerged as a branch of psychology dedicated to the study of the cognitive processes underlying verbal behavior. Most computational linguists (in particular with the dramatic increase of computer memory and speed) turned to the development of engineering tools to improve automatic processing of large text corpora.

A discussion of the reasons for this fragmentation, or its effect on the study of language in the last four decades, goes beyond the scope of this introduction (cf. Berwick and Weinberg 1984; Phillips 1996; Marantz 2005 for relevant discussion). Instead, we would like to express our conviction that 'the first decade of the 21st century should be a golden era for the cognitive neuroscience of language' (Marantz, 2005).

Work within theoretical linguistics has provided deep insights into the architecture of grammar. In computational linguistics over the last decade, a number of ambitious projects to computationally implement theoretical linguistic formalisms were brought to fruition (Harkema, 2001; Hockenmaier and Steedman, 2001; Shen and Joshi, 2005), reaffirming the connection between the two disciplines. A number of recent proposals have been made for computationally inspired human parsing models (Hale, 2001; Lewis and Vasishth, 2005) and psycholinguistically inspired computational models (Roark, 2001). Finally, the advent of in-vivo non-invasive neuroimaging in humans (fMRI, MEG, ERP, PET) has provided a rich, novel source of empirical data.

This thesis is part of a research program that builds on these new developments
in an attempt to realize the potential of an interdisciplinary, integrative view of language research. We make use of competence theory, combined with psycholinguistic measures and computational linguistic tools to model multiple dimensions of a sentence processing task. Our empirical data were collected using functional Magnetic Resonance Imaging technology.

As pointed out by Poeppel and Embick (2005) among others, data regarding brain function can be brought to bear on theoretical debates, but at the same time theoretical constructs and insights can serve to further our understanding of brain function. This latter prospect is particularly relevant in the case of language. First, lacking the appropriate animal models, we still know very little about the neural computation underlying symbolic cognition. Second (echoing Chomsky), language is arguably the cognitive domain for which we have the best computational model (grammar) to date. This fact allows us, in the domain of language, to investigate quantitatively complex internal representations and processes to an extent not currently possible in many other domains.

1.2 The use of fMRI in the study of language and sentence processing

For about 15 years now, functional Magnetic Resonance Imaging (fMRI) has allowed for non-invasive, real-time tracking of localized brain activation in humans. This technology has revolutionized the field and become the central source for data in cognitive neuroscience. The advent of non-invasive neuroimaging technology holds a great potential for the study of neural foundations of language. In the past, in the absence of a relevant animal model, the neural underpinning of the faculty of language could not be studied in-vivo and neurolinguistics was restricted, to a large degree, to broadly specified lesion studies.

Indeed, language has been a major topic for fMRI investigation since its inception, with hundreds of published studies to date. While fMRI has provided a rich source
of novel data, which has weighed on particular cognitive debates, in some linguistic domains (visual word processing is arguably the best example; cf. Jobard et al. 2003), it has been less fruitful in others, and in particular in the study of sentence processing and syntax. Two recent reviews of fMRI studies in the domain of syntax (Poeppel and Embick, 2005; Caplan, 2007a) paint a rather bleak picture of the state of the art. Poeppel and Embick even write that they ‘cannot conclude that major insights have been obtained concerning the structure of language or our understanding of the brain’ (p. 112).

Multiple factors contribute to this apparent failure (see the reviews cited above for a useful discussion). Below, we discuss, briefly, a number of these issues, that we find particularly problematic and that the design here represents an attempt to resolve.

1.2.1 The subtractive paradigm

The most dominant experimental paradigm in fMRI is the subtractive design. The subject is exposed to two conditions (A & B) which differ with respect to a well defined dimension of interest $\alpha$ (A contains $\alpha$ but B does not). Per-voxel$^1$ BOLD (blood-oxygen-level-dependent$^2$) response is measured for the two conditions, creating two functional (three dimensional) maps of the brain. The map of condition B is then subtracted from condition A. The difference is taken to be the neural activation that is causally related to the processing of $\alpha$.

This logic seems quite straightforward and the methodology can be (and has been) a very useful one. However, this approach cannot inform us what aspect(s) of $\alpha$ or of its processing is responsible for the observed activation. This problem is not unique to the study of syntax (cf. Caplan 2007b for an excellent discussion of fMRI experimental design) but it is particularly apparent in this domain. Consider the well

---

$^1$voxel is the three dimensional equivalent of a pixel

$^2$BOLD, which is a measure dependent on the relative amount of oxygenated hemoglobin in the blood and so indicates the amount of ‘fresh’ blood present in a region, is the standard measure used in fMRI research
studied linguistic phenomenon of passivization:

(1) Jim saw the ball.

(2) The ball was seen by Jim.

The standard view in early generative grammar was that the passive form (2) was derived from the active one (1) via a transformation. This basic insight has been maintained in current generative grammar, where the passive is considered to involve an additional movement (of the thematic object to the subject position). There are multiple evidence from typology, acquisition, neuropsychology and on line sentence processing measures, that passive sentences are more marked or difficult compared to their active counterparts.

At first glance, the subtractive paradigm appears appropriate, with (2) standing for A, and (1) standing for B. Linguistic theory provides us with a well defined α, namely the passive transformation, but this is where things become complicated. There are other possible candidates. Maybe it the fact that the active is more frequent that the passive? Maybe it is the added auxiliary or the particular morphology? Or maybe it is the unusual pragmatic effect of an inanimate noun phrase in a subject position or the mismatch between the subject role and the agent role? All these differences have been, in fact, proposed in the past to explain the observed markedness of the passive. The subtraction method cannot tell us which of these dimensions (or combination of) is responsible for the observed BOLD differences in fMRI experiments comparing passive and active (Yokoyama et al., 2006).

Put more generally, sentences are complex objects with multiple features and dimensions. It is virtually impossible to find a ‘minimal pair’ of two sentences which differ only along a single dimension or with respect to a single feature. This fact makes the interpretation of a subtraction design highly complicated.

Compounding this problem, in many of the published fMRI experiments on sentence processing using the subtraction design, the two conditions differ in ways that defy any coherent definition of α in terms of a linguistic theory. Friederici et al. (2000); Röder et al. (2002) compare sentences containing real open class words (con-
dition A) to sentences containing pseudo-words (jabberwocky, condition B). Though the authors define \( \alpha \) as *semantics*, we argue that this difference has no meaningful mapping onto a well defined linguistic notion. There is no theory of how jabberwocky sentences are processed (and the authors do not offer one) that can allow us to clearly define a possible \( \alpha \). Even a speculative definition of \( \alpha \) is problematic, since the notion of semantics these authors seem to have in mind collapses together lexical and compositional semantics that are kept distinct in any theoretical approach to language.

### 1.2.2 Deviationism

Grammaticality (or acceptability) judgments are the mainstay of work in theoretical linguistics. These are, in fact, psychological experiments on par with more ‘conventional’ psycholinguistic paradigms Marantz (2005). Linguists use ungrammatical or unacceptable strings to test possible hypotheses regarding the form of the mental grammar of the speaker. If a proposed grammar generates a string that is unacceptable by speakers, it is rejected (or modified). It is important to observe that linguists have no generative theory of (most of) the ungrammatical sentences used in these experiments\(^3\). In a related manner, linguists don’t take grammaticality judgments to be the functional purpose of the mental grammar in the same way that psycholinguists don’t take lexical decision to be the functional purpose of the mental lexicon.

Despite these truisms, ungrammatical strings and grammaticality judgments have been used extensively in fMRI investigations of sentence processing (cf. Kuperberg et al. 2006 for a recent example). Since there is no linguistic theory of ungrammaticality (putting aside footnote 3) it is not clear how to interpret brain activation associated with the processing of an ungrammatical string. It appears to us that the logic driving such experiments is that in the face of ungrammaticality the ‘syntax module’ is preferentially activated for some reason. The exact reason is often left implicit and, to our knowledge, none of these papers propose a theoretical framework

---

\(^3\)As pointed out by Marantz (2005) certain cases of ‘mild’ unacceptability as with certain weak island effects might represent derivable but, for some reason, dispreferred constructions.
for the processing of ungrammaticality⁴. In effect, Friederici et al. (2006) found that (in the same experiment) a grammaticality contrast produced activation in different regions compared to a syntactic complexity manipulation. The inferior frontal gyrus (Broca’s region, often associated with syntactic processing) was activated by the complexity manipulation but not the grammaticality contrast.

1.2.3 Meta-linguistic tasks and other issues of (un-)naturality

The issue of the use of grammaticality judgements in an imaging experiment is a part of a broader issue regarding the use of meta-linguistic tasks. Most fMRI language experiments make use of an on-line behavioral task either as an attention control (to make sure subjects are paying attention to the stimuli) or in order to probe a particular aspect of linguistic processing. Often, these tasks are meta-linguistic in nature, that is, subjects are asked to evaluate a linguistic dimension of the stimulus (plausibility, grammaticality, the presence of a particular word or phoneme, etc.)

The problem with meta-linguistic tasks as attention control is that most often the experimenters do not provide the relevant cognitive (not to mention neuro-cognitive) theory for that task. Since the tasks make explicit reference to language, it is probably the case that linguistic cognition is implicated in the task itself and so there is a real concern regarding interaction between the supporting task and the linguistic manipulation of interest. Caplan et al. (2007) investigated the interaction between the behavioral task and the linguistic dimension of interest. The linguistic contrast used was the well studied object (OR) versus subject relative (SR) clauses (Stromswold et al., 1996). The authors compared the fMRI activation for this contrast in the context of three different tasks; a sentence verification task (which requires passive reading of the sentence for a later semantic comparison to a probe sentence), a plausibility judgment and a non-word identification. The results for the linguistic contrast (OR-SR) differed dramatically across the three auxiliary tasks.

⁴The processing of speech errors and possibly certain stages of language acquisition, among other things, might require the ability to process ungrammatical strings, however, the above-mentioned literature is not concerned by these questions but instead makes claims about ‘normal’ processing.
Meta-linguistic tasks have also been used as the actual linguistic manipulation. In such experiments, the stimuli are kept identical across conditions but the task differs along the dimension of interest. The logic of this approach is similar to the logic used in paradigms in working memory and attention research (e.g. attending to or remembering the identity of an object versus its spatial position). Rogalsky and Hickok (2008), for a recent example, presented subjects with sentences, a subset of which contained either ‘syntactic’ violations (number agreement) or ‘semantic’ violations (incongruent animacy), in two conditions. In one condition, the subjects were instructed to identify syntactic violations and in the other, semantic violations. fMRI data was collected only from correct sentences, so the stimuli did not differ across conditions (all well formed sentences of English), only the task. The authors reason that when a subject is required to pay attention to subject-verb agreement their attention is shifted to the syntactic structure and when the subject is required to pay attention to animacy requirements their attention is shifted to the semantics of the sentence.

From the preceding discussion it should be quite clear that Rogalsky and Hickok’s task is not as straightforward to interpret as the authors suggest. What is the cognitive theory of an agreement violation detection task? In what way does it relate to the processing of agreement in ‘normal’ language use? Is it simply ‘more of the same’ as the authors seem to suggest? Or do subjects make use of an independent strategy? Maybe different subjects execute this (novel) task differently? Moreover, lacking a full understanding of the two tasks, it is not unlikely that they differ on additional dimensions beyond the difference in linguistic domain. Looking at the activation difference across conditions, how are we to distinguish between these activation differences that are due to the different linguistic dimension and those that are due to other, orthogonal, differences between the tasks?

The problematic issue of meta-linguistic tasks is part of an even broader concern in neurolinguistic (and psycholinguistic) experiments, namely that of ecological validity. Subjects are presented with isolated sentences (or even words) in a situation which has none of the features of a normal communicative context. The stimulus sentences often
contain definite descriptions and other presuppositional items\(^5\) without a supporting context. The subjects are then required to perform a (meta-linguistic) task on these linguistic objects which, again, has little to do with what people usually 'do' with language. While normal, cooperative, adults are able to overcome these pragmatic challenges and perform the required task, the effect of these non-ecological factors on brain activation is rarely brought into consideration. In addition, young children and subjects belonging to special populations (e.g. autistic individuals, lesion patients, etc.) might not be as successful in compensating for these factors.

### 1.3 A triple C paradigm

In this thesis we used a novel fMRI paradigm inspired by an on-going research project at NYU (Brennan et al., 2007). Our paradigm makes use of a parametric design (1.3.1) to model, word-by-word, the processes underlying sentence comprehension. Our model of the cognitive task makes a novel use of both theoretically and computationally derived measures (1.3.2), instantiating a triple C approach to neurolinguistics. Finally, the paradigm makes use of a passive listening task and structured narratives, in an attempt to maximize naturalness (1.3.3).

#### 1.3.1 A parametric design

Caplan (2007b) distinguishes between *qualitative* contrasts and *parametric* contrasts. In a qualitative contrast one of the conditions (condition A in 1.2.1) requires a processing stage or an appeal to a representational capacity (\(\alpha\)) of a type not required by the contrastive condition (condition B). In a parametric contrast the two conditions are taken to require the cognitive process of interest \(\alpha\) but A requires more of it (by some metric) compared to B. Consider Ben-Shachar et al. (2004) and Santi and Grodzinsky (2007). In both experiments Grodzinsky and colleagues were interested in the neural underpinning of syntactic movement. In the earlier experiment they

used a *qualitative* contrast, comparing sentences containing syntactic movement with sentences containing no such movement. Brain regions that appeared active in this contrast were interpreted as supporting movement, taken to represent a qualitatively distinct aspect of grammar.

In the 2007 paper, the authors made use of a *parametric* contrast\(^6\). In both conditions the sentences contained movement but the distance between the antecedent and the gap in condition A was greater than in condition B. \(\alpha\), in this case, was the working memory component required by the long distance dependency between the antecedent and the gap. The distance was the metric. Condition A required more of the relevant working memory than condition B.

Parametric designs are particularly appealing in the study of language (and in particular sentence processing). First, as discussed briefly above, it is not always obvious how to create a condition B that does not contain the cognitive function of interest. What does it mean, for example, to have a sentence processing condition that does not make appeal to semantic interpretation (Friederici et al., 2000)?

Second, as can be observed in the example above from Grodzinsky’s work, a qualitative contrast require the two conditions to contain different syntactic constructions. As was discussed earlier with respect to the passive construction, it is not always obvious how to isolate a single dimension of interest when comparing two constructions since they can differ on multiple dimensions. Another example of such confound is the subject relative/object-relative contrast used by Caplan and colleagues in a number of reported experiments (Stromswold et al., 1996; Caplan et al., 2003, 2007):

\[(3)\] The boy spilled the juice that stained the rug.

\[(4)\] The juice that the boy spilled stained the rug.

Even if we restrict ourselves to purely syntactic dimensions, the object relative sentence (4) differs from the subject relative on (at least) two accounts. First, the object relative contains a long non-local dependency between *juice* and *spilled*. Second, the

\[^{6}\text{The full experiment contained also a qualitative contrast between binding and movement. We restrict our attention here to the parametric manipulation}\]
object relative sentence contains a center embedding of the relative clause, while the subject relative sentence is completely right branching. In Santi & Grodzinsky's parametric design, the construction was kept constant, keeping the two conditions more similar.

A third advantage of parametric contrasts stems from their high compatibility with regression designs. Qualitative contrasts are, by nature, binary. Parametric contrasts are not constrained in this way. While some dimensions of linguistic objects are binary (for example, the voicing distinction in phonology), many others are continuous and can take many values along the relevant metric (consider length or frequency, for example). Though many experiments reduce these continuous variables into binned contrasts (short versus long, high frequency versus long frequency), one can alternatively make use of the full power of the linear model\(^7\), using multiple regression in such cases.

Multiple regression, in turn, has the power to disentangle the contribution of multiple independent variables to the observed dependent variable. As has been highlighted above, this feature is very useful for the study of syntax, given the highly complex nature of the stimuli and the multiple dimensions that come into play in sentence processing. Using multiple regression eliminates the notion of conditions. Each stimuli represents a point in a multi-dimensional space (defined by the different linguistic variables). Brain data from individual ‘events’ are not averaged (a ‘single trial’ design). This single trial aspect has been critical for this experiment where each word in the narrative has been modeled an event or a trial.

1.3.2 Theoretical and computational measures

Our multiple regression parametric design included variables which modeled numerous linguistic dimensions. Some of the variables modeled basic features of the input (for example, the length of each word). Other variables quantified distributional properties of the individual words (e.g. frequency). Yet other variables represent hypotheses

\(^7\)t-test and ANOVA are just special cases of the general linear model
regarding the internal state of the listener and are related to the incoming stimuli only indirectly.

The results reported in this thesis stem largely from this last category of variables. As we discussed above, our ability to quantify complex internal states and not only direct properties of the stimuli is due to the existence of a sophisticated generative competence models in theoretical linguistics and sophisticated computational models of parsing in natural language processing.

Non-local dependencies, the topic of chapter 5, are a prime case of a theoretical construct. The different aspects of these dependencies (the relevant syntactic operations, working memory resources, etc.), which will be quantified here are not a property of the stimuli but are derived from theoretical models of syntax and sentence parsing. The use of such theoretical notions is not novel but their use in a parametric design has been limited.

To our knowledge, the use of the internal state of a computationally implemented parser to generate predictions regarding brain activation is novel. It is important to note that the use of a computational tool here goes beyond the use of other computational tools in recent fMRI experiments. A number of recent experiments correlated the outputs of machine learning algorithms applied to a set of stimuli with brain data acquired while subjects viewed the same stimuli (cf. Mitchell et al. 2008 for an application of this method to language). These experiments use computational tools to extract complex relations between stimuli or to identify otherwise hard to see patterns.

Here, we used an incremental parser and used the ‘partial knowledge’ of the parser at every word regarding the final output as a model of a human listener. In other words, we took the on-line computations performed by the parser to stand for the computations performed by the human subjects in the scanner. While the literature mentioned above makes use of computational models, here we model computation.
1.3.3 A naturalistic task and context

In the current state of the art, fMRI experiments impose severe limits on 'naturalness'. Subjects are lying, motionless, in a narrow noisy tube, interacting with a computer via buttons and a headphone. Many aspects of a normal communicative situation cannot be replicated in these conditions. Modulo these obvious limitations, the experimental design that we used here was intended to address many of the issues raised in 1.2.3.

We used auditory presentation, which avoids the problem of word-by-word presentation used by most fMRI experiments of sentence reading. Natural reading makes use of repeated scans and back saccades, as well as para-foveal information. All these are not available in word-by-word reading and so reading in the scanner is quite a novel task. Auditory presentation is not limited in that way.

Subjects listened to (on average) five minute long narratives rather than single sentences or short passages. Each of the narratives was written in a recognizable prose style (a journal article, historical account, etc.) to enhance the naturalness of the text and to draw attention away from the particularly complex syntactic structure. The embedding of the target structures in a rich linguistic context supported the use of presuppositional items and avoided the coherence issues discussed earlier.

The attention control task used here was an off-line comprehension test, delivered to the subjects after each of the narratives. It is plausible that, since the subjects know they will be asked questions later, they might be using special strategies to better memorize the incoming narrative. One could argue that this additional strategy is open to the same criticism we raised regarding the often used on-line meta-linguistic tasks. This is correct, however, we would argue that such a task is (a) ecologically valid, since text comprehension and retention is a skill often required in the 'real world' and (b) minimally disruptive, since it eliminates the decision and decision related effects (emotional and other) from the scanning period.
1.4 The structure of the thesis

The next section in the introduction (1.5) contains a technical description of the experimental design, including details regarding the subjects, stimuli, linguistic measures, fMRI preprocessing and statistical analysis. The final section of the introduction contains the results of the word rate (1.7.1) and word length (1.7.2) contrasts.

Chapter 2 raises the issue of the role of prediction and statistical knowledge during sentence processing. Data is reported from contrasts reflecting the frequency and predictability of lexical items as well as syntactic complexity and surprisal.

Chapter 3 focuses on the issue of structural ambiguity and its effect on processing during the ambiguous region and at the point of resolution. This chapter reports the results of contrasts that make use of a novel measure reflecting word-by-word entropy (uncertainty).

In chapter 4 we look at the issue of Theory of Mind, or the capacity to represent other people’s beliefs or intentions. Using a sentence-level measure, we demonstrate that our naturalistic design replicates known facts regarding the neural network subserving this capacity. We then present data from more fine-grained word level contrasts that tease apart different aspects of Theory of Mind (namely beliefs and intentions) and its linguistic representation (modality).

Non-local dependencies are the topic of chapter 5. A sentence level parametric measure of dislocation provides a general outline of the brain regions engaged in the processing of such dependencies. We make use of the fine-grained resolution of our word level model to distinguish the effects of retrieval from those of maintenance. We also compare the neural imprint of two types of non-local dependencies distinguished in the theoretical literature.

In chapter 6, the final chapter, we outline a neural model of grammatical competence. This model serves as a syntheses of the results reported in the preceding chapters in the context of the existing functional and anatomical imaging literature.
1.5 Methods

1.5.1 Participants

Subjects were recruited from MIT and the surrounding community. 23 volunteers took part in the experiment. 3 subjects were excluded for low scores on the comprehension component. Data from two other subjects was lost due to technical malfunction. All remaining 18 subjects (12 females) were healthy, right handed, native speakers of American English with normal hearing and language abilities. Mean age of subjects was 23 (range 18-31). Subjects received monetary compensation for their participation that included a small success based bonus to encourage them to pay attention to the narratives and questions.

1.5.2 Procedure

The study was conducted at the Athinoula A. Martinos Imaging Center at McGovern Institute for Brain Research, MIT. Each experiment was administered over two sessions. During the first session subjects were debriefed and provided an informed consent. Before entering the scanner, participants performed a practice trial, administered using a laptop computer running PsyScope X (Cohen et al., 1993). Once in the scanner, subjects went through 4 experimental trials. Each trial consisted of a passive listening to a narrative, followed by 12 auditory comprehension questions. Each question was followed by a visually presented answer to which the subject was required to respond with a true/false judgment. The presentation order of the four narratives was randomized across subjects.

PsyScope was used for all stimulus presentations well as the recording of subjects’ responses. The auditory narratives and comprehension questions were delivered using electro-static headphones. An Hitachi (CP-X1200 series) projector was used for all visual presentations with the image projected through a wave guide and displayed on a rear projection screen (Da-Lite). Immediately following the scanning session subjects completed a syntax memory forced-choice test (to be described below) followed by
the North American Adult Reading Test (NAART, Uttl 2002). During the second experimental session subjects completed a battery of working memory tests (Waters and Caplan, 2003).

1.5.3 Materials

The narratives

The four narratives were written specifically for the purpose of this experiment. A professional writer was given instructions regarding the desired syntactic structure manipulations (varying types and lengths of non-local dependencies, varying embedding depth, etc.) and produced a first draft. Each narrative was written in a different style (journalistic report, historical account, fiction). This stylistic decision was taken for two reasons. First, we reasoned that different narrative styles will make certain syntactic structures more or less natural. Second, this choice served as an implicit orthogonalization of the syntactic measures from the specific context or more global narrative structure. Both decisions were based on intuition rather than a quantified result. The author reviewed the drafts before a first revision by the same writer. In a second stage, the narratives were revised by native English speaking linguists. The linguists were instructed to add complex syntactic structure such as raising constructions and additional relative clauses. The narratives were then recorded, spoken by a native American English speaker (all narratives were read by the same speaker), in a sound attenuating booth, using Amadeus software running on a Macintosh computer. The resulting audio files were edited using Praat software (Boersma and Weenink, 2008). Praat was also used to manually segment the audio files at three different granularities (sentences, words, syllables), allowing us to identify the onsets, offsets and durations of the relevant linguistic events. Other acoustic data was also extracted (e.g. F0) but was not used for the analyses presented here. The total length of the 4 narratives was just under 22 minutes with an average length of 5.5 minutes. The full text of the four narratives can be found in B.
Linguistic coding

Multiple measures were calculated per word and per sentence using manual coding as well as automated tools, namely the Roark parser (Roark, 2001), CPIDR (Brown et al., 2008) and the English Lexicon database (Balota et al., 2007). Before reviewing the relevant measures we will briefly introduce the automated tools mentioned above.

- Roark’s parser: Roark’s parser makes use of a lexicalized probabilistic context-free grammar and an incremental top-down strategy (it was developed with automatic speech recognition in mind, which requires an incremental parser). The parser makes use of two important transforms. The first is a selective left-corner transform by which recursive rules (such as NP→ NP PP) are transformed to sequences of right-branching rules (NP→DT NN NP-NP and NP-NP→PP). This is useful since left recursive structures are memory intensive for a top-down parser (Abney and Johnson, 1991; Resnik, 1992). The second transform, a left-factorization, essentially turns all rules into binary productions where the second term on the right hand side is an under-specified non-terminal. For example, the rule VP→ V NP PP becomes VP→ V VP/S where VP/S stands for all possible continuations that, combined with V would produce a VP. The under-specification prevents (when parsing the verb in the example here) an explicit computation and storage of all possible expansions of VP or alternatively an early prediction of the correct VP structure. Instead, the new ‘slash’ non-terminal (which resembles a categorial grammar category) collapses all possible continuations into one rule. This allows for the parser to await more information before selecting the correct VP expansion rule. A second advantage of this transform, from our perspective, is that it turns the often flat Penn tree-bank representations into right branching binary trees which are closer to current view of English syntax.

One important advantage of a top-down strategy (coupled with the transforms discussed above) for incremental probabilistic parsing is that it produces rooted derivations for every sub-string (or prefix). This feature allows the parser to
generate probabilities for each prefix directly from the grammar (by summing the probabilities of all possible derivations that would generate that prefix). In addition, the parser can make use of the entire left context to condition the probability of the specific rule. Roark's parser makes a selective, linguistically informed, use of the left context (conferring particular importance, for example, to c-commanding, local, terminals). Finally, Roark’s parser employs a parallel architecture to parsing and so at every given moment multiple candidate parses (each with a specific probability) are kept in the parser’s memory. This list is referred to as the beam of the parser. The beam is thresholded by probability so certain possible, but highly unlikely parses, are discarded.

While many features of Roark’s parser have been inspired by linguistics and psycholinguistics (Brian Roark p.c.) it is important to bear in mind that it is primarily an engineering tool and that the Penn tree-bank annotation of the Brown corpus (used to train the parser for the experiment here) is quite different from current assumptions regarding the syntactic structure of English. Even more important is the fact that this parser (as most other broad coverage parsers) is based on a context-free grammar, which has been shown to be insufficient to capture the full complexity of natural language (Shieber, 1985). As a consequence, the use here of the Roark parser states to model human parsing must be understood as a (far from perfect) idealization. In future work with Brian Roark we intend to use additional tree transforms to make the grammar used by the parser more similar to the current state of the art in theoretical linguistics.

- **CPIDR**: CPIDR (Computerized Propositional Idea Density Rater, Brown et al. 2008) generates Propositional Idea Density (PID) measures for a given text. It is assumed that propositions are the units involved in understanding and remembering texts. Following Kintsch (1974), the main verb and all its arguments (subject, object, indirect object, etc.) are taken to be one proposition. Additional descriptive elements such as adjectives, adverbs, and qualifier phrases are
additional propositions. PID is the number of expressed propositions divided by the number of words. This measure (which was computed manually in the past) has been shown to correlate with readability, writing quality, and memory. In addition it has been useful in the prediction of Alzheimer’s disease (cf. Brown et al. 2008 for relevant references). CPIDR first tags the input text using the Penn treebank POS (part of speech) conventions. At the second stage a number of readjustment rules are used. For example, sequences of ‘either...or’ as well as auxiliary sequences (may have been singing) are counted as a single proposition. The authors demonstrate that ratings produced by CPIDR for 80 speech samples were highly correlated with PID rates produced by hand (r=0.9693).

- The English Lexicon Project: The ELP is an extensive database of 40,000 words collected as a part of a collaboration across 6 universities Balota et al. (2007). The database is unique in that it combines descriptive characteristics of the words (length, frequency, etc.) with normative data for speeded naming and lexical decision (across 1200 subjects).

Below we provide a summary of the measures used in the analyses reported here. Most measures will be discussed again in the relevant chapters.

- Word length: As discussed above, word length (in ms) was determined by manual inspection of the auditory file. A sentence level measure was computed by averaging the length of all open class words.

- Lexical class: We divided words into open and closed class categories. There is no consensus in the literature regarding the putative boundary between the two categories. The words were first automatically tagged and the tags were used for classification. All nouns, verbs (except auxiliaries) adjectives and adverbs were considered open class items. Determiners, connectives (conjunctions, disjunctions, etc.), auxiliary verbs and pronouns were considered closed class items. This classification was used as a parametric modulator in one of the models (the Word model), as a basis for distinguishing two conditions (in the Word-Class model) and to produce class specific averages in the Sentence model.
• Log Frequency: We used disambiguated word lists from our narratives to query the ELP for (log) frequency estimates for each word based on the HAL corpus (Lund and Burgess, 1996). The sentence level frequency measure was calculated as the averaged log frequency value across all open class words in the sentence.

• Number of derivation steps: This measure was computed directly from the internal state of Roark’s parser at each word. It is the number of steps (or rules) used during the integration of a given word into the incrementally constructed syntactic structure(s). It reflects, among other things, the local structural complexity of the sentence.

• Surprisal: Lexical and syntactic surprisal measures for each word were computed on the base of the prefix probabilities generated by the Roark parser. Sentence level surprisal (or entropy) was computed as the average open class lexical surprisal and average syntactic surprisal across all words.

• Ambiguity: An ambiguity and a disambiguation measures were produced on the base of the probability distribution over the beam of the parser at each word.

• Theory of mind (TOM): TOM measures were computed both at the sentence level (distinguishing sentences about TOM from other sentences) and at the single word level. At the word level three different binary (0 1) measures were computed. One measure was sensitive to belief state references, one to references to intentions and third to all other modal expressions. All these measures were computed manually by the author and another linguist based on linguistic intuitions in the absence of any general definitions.

• Non-local dependencies: Since the Roark parser was not designed to handle non-local dependencies these measures were calculated manually. A trained linguist tagged each instance of a word preceding a gap (retrieval). Gaps were divided into two types (so called A and A-bar), following a general consensus in the field. The interval between an explicitly marked antecedent (e.g. a Wh word) or any other word indicating the presence of an A-bar dependency (e.g. a
relative pronoun) and the associated gap was also tagged. In addition, following Gibson (2000) each A-bar gap was tagged with the number of discourse referents intervening between the antecedent and itself. A sentence level measure was computed by normalizing the number of gaps by the total number of words.

- Propositional Idea density: PID was calculated per sentence using CPIDR.

**Behavioral controls**

- Comprehension questions: 12 comprehension questions were prepared for each narrative. Half the questions were polar (yes/no) questions and the other half constituent (wh) questions. The questions were recorded by the same speaker that read out the narratives. In order to make the response method to the two questions similar, one written answer was prepared for each question, which was false for half of the questions and true for the other half.

- Post scanner syntax memory test: A written list of sentences (10 for each narrative) was prepared. Half of the sentences were actual sentences from the narrative while the other half were syntactic transformations of sentences that appeared in the narrative (for example, an active voice sentence from the narrative could appear in the passive voice at test or vice-versa). This test was administered in a forced choice design.

**1.5.4 Data acquisition**

Functional and structural Magnetic Resonance Imaging data was acquired in a 3T Siemens Tim Trio scanner (Martinos Imaging Center at the McGovern Institute for Brain Research, MIT) with a 12 channel head coil. Each session started with a whole head structural (T1) MPRAGE sequence (1-mm*1-mm*1-mm resolution). Functional data for the four narratives were acquired using a whole head 3D-PACE EPI sequence with TR of 2 seconds and 3-mm isotropic voxel size (31 slices). PACE, or Prospective Acquisition Correction, is an inline motion correction techniques which adjusts the acquisition gradients so to minimize any effects of head movement between one
functional volume and the next. Apart from the usual first 4 volumes which are automatically discarded (to avoid large intensity drop), the first 9 saved volumes consisted a rest period where the subject was simply looking at a fixation point. The narrative was triggered by the the 10th TR. Once the narrative ended, another 10 volumes were acquired to measure the decline of the hemodynamic response function (with the subject, again, looking at the fixation point. The comprehension question section was also monitored with identical acquisition protocol but that data is not analyzed here.

1.5.5 Data preprocessing

Images were obtained from the scanner in DICOM file format then converted to the Nifti file format using mri_convert (part of FreeSurfer; Dale et al. 1999). SPM5 (Wellcome Trust Centre for Neuroimaging) utilities were used for motion correction and extraction of motion parameters, and spatial normalization to standard MNI space. All images were smoothed by convolution with an isotropic, three-dimensional Gaussian kernel of full width at half maximum of 8mm. Images were examined for outliers using the Artifact Detection Tools available as part of the Brain Imaging Tools (http://web.mit.edu/swg/software.htm). Outlier timepoints were identified as volumes with a global mean that exceeded 2.8 standard deviations from the mean or with estimated subject motion greater than 0.5 mm of translation or 0.01 radians of rotation in any dimension. Outlier volumes were excluded from later analysis by including a dedicated nuisance regressor for each in the general linear model. All functional images were visually inspected.

1.5.6 Data analysis

Generalities

All statistical analyses were done using SPM5’s implementation of the Generalized Linear Model (GLM). Three different first-level (single subject) models were created to analyze the data set. In each model regressors of interest were introduced either
as independent conditions or parametric modulations of an existing condition. All
three models were augmented with a session specific constant (to normalize the data
across the 4 separate runs), movement regressors and a regressor for each artifact
image (if any were detected by ART). All regressors of interest were then convolved
with SPM's canonical hemodynamic response function (HRF). This convolution is
a commonly employed strategy to relate expected neural activity to the observed
BOLD response that reflects hemodynamic, rather than neural, dynamics. Previous
experimental work has established a 'canonical' hemodynamic (or BOLD) response to
a neural event that peaks at about 5 seconds after the neural event itself and slowly
decays (returning to baseline at around 25 seconds). By convolving the predictors
with an idealized HRF, we turn the neural predictors into BOLD predictors.

The coefficients (or relative contribution of each regressor to the linear model of the
observed data) of each model were then estimated for each voxel (minimal volume
unit) and contrast maps computed. A contrast represents a linear combination of
coefficient estimates. A contrast can represent the effect of a single regressor, a
combination of regressors or a difference between two regressors. A contrast map
reports the contrast value for all voxels in the volume (or brain). For each contrast,
the relevant maps of all subjects were combined and t-tests were used to estimate the
significance of the effect at the group level (again, per voxel). The resulting statistical
parametric maps (SPM), containing the t value for each voxel for each contrast were
thresholded at either 0.001 or 0.005, uncorrected for multiple comparisons, with a 20
voxel cluster size threshold.

For the purpose of additional anatomical precision and easier presentation, these
SPMs were overlaid on a surface based representation of the MNI canonical brain using
the SPM Surfrend toolbox (written by I. Kahn; http://spmsurfrend.sourceforge.net).
The surfaces were then rendered using FreeSurfer (CorTechs Labs, Inc., Charlestown,
MA) (Dale et al., 1999; Fischl et al., 1999). The results are reported both visually and
in tables containing the coordinates of all cluster activation peaks, or local maxima,
more than 8 mm apart. Each table reports the x,y,z coordinates in MNI (Montreal
Neurological Institute) space of the local peak, its t value and general anatomical lo-
cation as defined, manually, using (for cortical activation) the automatic anatomical annotation produced by Freesurfer and (for sub cortical activation) the Talairach Atlas Label as determined by xjView (http://people.hnl.bcm.tmc.edu/cuixu/xjView/). For easier reading, the peaks were grouped and ordered by anatomical location rather than activation strength.

The three models

One advantage of the approach here is that we can model the data at different granularities. For the purpose of the work presented here two granularities were selected; namely the word and the sentence. The sentence model had a single condition (‘sentence’) which was constructed as a block design with each sentence in the narrative constituting a single block. In addition, the model included 9 parametric modulators, namely: Averaged lexical surprisal, averaged syntactic surprisal, averaged ambiguity, averaged log frequency, averaged word length, averaged derivation steps, Theory of Mind, Propositional Idea Density and normalized gap value.

For the word level we produced two different event related models (Word and Word-class). The reason for this duality is the interaction of length, word class and frequency/lexical surprisal. Closed class words were (unsurprisingly) shorter, much more frequent, and less surprising compared to open class words. In addition, the variance of these measures was much lower for the closed class words. As a consequence, the surprisal/frequency effect would have been potentially confounded with the class difference. In order to avoid this confound we put together a word-class model with a separate ‘open class’ and ‘closed class’ conditions. In the ‘open class’ condition, the onset of every open class word was treated as an event, and similarly for the closed class words in the ‘closed class’ condition. In addition, each condition was parametrized using the following measures: Lexical surprisal, log frequency, syntactic surprisal, number of derivation steps, A gap, A-bar gap, word length, ambiguity and reanalysis (or disambiguation cost). A third condition, ‘maintenance’ was specified by blocks, each such block spanning the period of a-bar maintenance (or period between the antecedent, or filler, and its associated gap).
The second word level model (Word) had a single ‘word’ cognition that did not distinguish between word classes. This model was better suited to look at the effect of measures that were not affected by the open/closed class distinction (such as ambiguity, non-local dependencies and theory of mind. This model included 13 parametric modulators: Lexical surprisal, log frequency, syntactic surprisal, number of derivation steps, A gap, A-bar gap, word length, ambiguity, reanalysis, belief states, intentions, modality and lexical class. Similarly to the Word-class model, the Word model also included a ‘maintenance’ condition.

Co-linearities in the design

One potential source for complication is the fact that different dimensions of the linguistic object (be it a sentence, a word or a sublexical unit) are often not independent. The inverse relation between word frequency and length (Zipf, 1935) is one well know example. In a typical psycholinguistic experiment using a factorial design, the experimenter would select or construct the stimuli so to control for possible confounds (for example, if the experimenter is comparing high to low frequency word, she will make sure to select words so that there will not be a length difference across the conditions). In our experiment, which has more naturalistic stimuli, this type of control is unfeasible. Instead we make use of multiple regression and model simultaneously the different dimensions of the task. In a sense, we allow the different regressors to compete with each other in explaining the variance in the data. This approach is conservative since any shared variance will not be uniquely attributed to any single regressor. If two regressors are highly correlated than we might not be able to produce an estimate of the coefficients of these regressors. However, even in the presence of mild correlation, estimation is possible, though, of course, it will be less accurate (or will have greater variance).

Most fMRI experiments using the GLM approach do not encounter this problem since the experiments are designed such that the regressors of interest are temporally orthogonal to start with (e.g. whenever condition A is 1, condition B is 0 and vis versa, or in other words, each event in the model belongs to only one condition).
Since we are modeling different dimensions of the same events and moreover the dimensions are not independent, co-linearities are expected. SPM5 deals with co-linearities between the main condition and the parametric modulators by recursively orthogonalizing the parametric modulators with respect to the main condition. This approach is fine when one deals with a single parametric modulator or if one is not interested in the effect of the parametric modulators themselves. However, we have multiple parametric modulators and using SPM5’s orthogonalization we would have forced an a-priori order of importance on the modulators (since each additional modulator is orthogonalized with respect to all earlier ones). Following the advice of Eric Zarahn (p.c.) and Alfonso Nieto-Castanon (p.c.) we have decided instead to mean correct all the parametric modulators. This manipulation eliminates the co-linearity between the main condition and the modulators but preserves any co-linearities between the modulators themselves. An alternative solution could have been to apply a Principle Component Analysis (PCA) or a similar method to the linguistic regressors and produce new, orthogonal, components. This approach has been used by Hauk et al. (2006, 2008) to resolve similar colinearities across multiple word level features in a lexical decision task. The problem with this approach is that the interpretation of the resulting components is not straightforward.

Fortunately, colinearity (or correlation) across the different regressors was overall quite low (with r mostly < 0.1). The major exceptions to this are the surprisal measures. Lexical surprisal correlated (negatively) with log frequency and (positively) with word length. Syntactic surprisal correlated positively with number of derivation steps. We return to this issue in (2.6.1). In, we provide the correlation matrices of the three models (only for open class words for the Word Class model):

**The hemodynamic response function**

In our word level models, word onsets are conceptualized as (neural) events. These neural events are then modulated by the different parameters discussed above. However, fMRI, contrary to electrophysiological measures such as MEG or ERP, does not have the temporal resolution (< 1 second) that would allow a direct observation of
the neural response for each word (typically 3 words per second). Instead, the dependent measure quantified by fMRI (the hemodynamic response) can be thought of as an integrator over a larger temporal window (> 5 seconds) with a sampling rate (in this experiment) of 2 seconds.

In a recent narrative reading experiment with a similar approach to the one de-
scribed here (Yarkoni et al., 2008), the authors chose to down-sample the values of their regressors (which reflected presentation rate of about 3 words per second) by averaging the values of each repressor within each TR (2 second) intervals. This choice does bring the predictor to a similar resolution as the actual sampling rate, however, it loses the more fine temporal characterization of the stimuli (and the pu-
tative neural response) which have been shown to be reflected in the BOLD signal despite the resolution limit. In such an approach, the effect of a linguistic measure at time X is modeled to have the same contribution to the signal at X+5 seconds as the same measure at time X+2 seconds (since the the measures are averaged together in 2 second bins). We find this consequence to be unfortunate. Instead, we preserve
the original temporal structure of the linguistic input by turning every word into an event. Naturally, the HRF convolution smoothes the original regressors but does not eliminate the sub-resolution temporal differences. In a sense, this approach can be thought as time-weighted averaging of the linguistic measure at each TR.

This approach has a second, related, positive feature. In the experiment here, words were presented auditorally in a natural (and so non-static) rate. As a consequence (and in contrast to Yarkoni et al. 2008 where presentation rate was fixed), each TR might contain a different number of words (due to pauses, speaking rate, word length, etc.). Since each word is modeled as an event, which is then linearly convolved with an HRF, the more words appear within a given time period, the larger the hemodynamic predictor is. This fact turns the ‘word’ event in our design into an effective ‘word rate’ measure. In other words, due to temporal resolution we do not expect the BOLD signal to exhibit an independent peak to each word onset but instead to reflect the density (or rate) of lexical input in the associated time period. We used this ‘word rate’ measure as a basic probe for linguistic processing. In other words, instead of using a subtraction measure (listening to narrative - listening to scanner noise) we used the internal dynamics of the narrative itself (the changing presentation rate). We predicted that higher word rate would produce more activation in the language processing network. The results of this measure, as well as that of word length will be presented next.

1.6 Behavioral results

Three of the subjects scored below 60% correct on the comprehension questions and were excluded from further analysis. The remaining 18 subjects scored on average 87% on the comprehension questions. Performance on the post-scanner memory for syntactic structures did not differ from chance for all subjects.
1.7 FMRI results

In this section we report the results of two contrasts, namely word rate from the *Word* model and open class word length from the *Word-class* model. All other results will be reported in the relevant chapters.

1.7.1 Word rate

Peaks of positive and negative correlation with word rate ($P < 0.001$ uncorrected, cluster threshold =20) are reported in tables (1.1,1.2):

<table>
<thead>
<tr>
<th>Region</th>
<th>Location</th>
<th>MNI coordinates (x,y,z)</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>L temporal</td>
<td>superior temporal sulcus</td>
<td>-58 -16 -2</td>
<td>13.985</td>
</tr>
<tr>
<td></td>
<td>middle temporal gyrus</td>
<td>-62 -36 4</td>
<td>13.077</td>
</tr>
<tr>
<td></td>
<td>superior temporal gyrus</td>
<td>-56 -8 -6</td>
<td>12.2987</td>
</tr>
<tr>
<td></td>
<td>rectus gyrus</td>
<td>-4 50 -16</td>
<td>7.1231</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 40 -22</td>
<td>5.856</td>
</tr>
<tr>
<td></td>
<td>precentral gyrus</td>
<td>-48 -4 56</td>
<td>5.8565</td>
</tr>
<tr>
<td></td>
<td>middle frontal gyrus</td>
<td>-48 4 52</td>
<td>5.2591</td>
</tr>
<tr>
<td></td>
<td>precentral gyrus</td>
<td>-40 2 44</td>
<td>4.9847</td>
</tr>
<tr>
<td></td>
<td>medial superior frontal gyrus</td>
<td>-4 14 62</td>
<td>5.786</td>
</tr>
<tr>
<td></td>
<td>orbital part</td>
<td>-52 24 -8</td>
<td>5.7303</td>
</tr>
<tr>
<td></td>
<td>opercular part</td>
<td>-60 16 24</td>
<td>5.6989</td>
</tr>
<tr>
<td></td>
<td>opercular part</td>
<td>-48 16 24</td>
<td>5.6768</td>
</tr>
<tr>
<td></td>
<td>superior frontal gyrus</td>
<td>-10 38 52</td>
<td>5.5378</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-8 54 42</td>
<td>5.0534</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-6 56 34</td>
<td>4.6458</td>
</tr>
<tr>
<td></td>
<td>anterior superior frontal gyrus</td>
<td>-12 60 18</td>
<td>4.5667</td>
</tr>
<tr>
<td></td>
<td>anterior hippocampus</td>
<td>-24 -12 -16</td>
<td>4.8479</td>
</tr>
<tr>
<td></td>
<td>posterior hippocampus</td>
<td>-14 -30 -5</td>
<td>4.7355</td>
</tr>
<tr>
<td></td>
<td>superior temporal gyrus</td>
<td>60 -8 -4</td>
<td>15.3785</td>
</tr>
<tr>
<td></td>
<td>posterior superior temporal sulcus</td>
<td>44 -54 20</td>
<td>5.2335</td>
</tr>
<tr>
<td>R frontal</td>
<td>orbital part</td>
<td>46 30 -12</td>
<td>4.336</td>
</tr>
<tr>
<td>R sub-cortical</td>
<td>cerebellum</td>
<td>20 -80 -22</td>
<td>5.5113</td>
</tr>
<tr>
<td>L sub-cortical</td>
<td>anterior superior frontal gyrus</td>
<td>-12 60 18</td>
<td>4.5667</td>
</tr>
<tr>
<td></td>
<td>anterior hippocampus</td>
<td>-24 -12 -16</td>
<td>4.8479</td>
</tr>
<tr>
<td></td>
<td>posterior hippocampus</td>
<td>-14 -30 -5</td>
<td>4.7355</td>
</tr>
<tr>
<td></td>
<td>superior temporal gyrus</td>
<td>60 -8 -4</td>
<td>15.3785</td>
</tr>
<tr>
<td></td>
<td>posterior superior temporal sulcus</td>
<td>44 -54 20</td>
<td>5.2335</td>
</tr>
<tr>
<td></td>
<td>orbital part</td>
<td>46 30 -12</td>
<td>4.336</td>
</tr>
<tr>
<td></td>
<td>cerebellum</td>
<td>20 -80 -22</td>
<td>5.5113</td>
</tr>
</tbody>
</table>
Table 1.2: Word rate results (negative)

<table>
<thead>
<tr>
<th>Region</th>
<th>Location</th>
<th>MNI coordinates (x,y,z)</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>L temporal</td>
<td>posterior superior temporal sulcus</td>
<td>56 -46 38</td>
<td>7.1522</td>
</tr>
<tr>
<td>L frontal</td>
<td>middle frontal sulcus</td>
<td>-32 52 28</td>
<td>7.818</td>
</tr>
<tr>
<td>L parietal</td>
<td>middle frontal gyrus</td>
<td>-30 34 44</td>
<td>5.8645</td>
</tr>
<tr>
<td>L parietal</td>
<td>superior parietal gyrus</td>
<td>-46 -56 56</td>
<td>5.6061</td>
</tr>
<tr>
<td>L parietal</td>
<td>Inferior parietal</td>
<td>-34 -42 34</td>
<td>5.6008</td>
</tr>
<tr>
<td>L parietal</td>
<td>medial superior parietal gyrus</td>
<td>-6 -76 44</td>
<td>5.2607</td>
</tr>
<tr>
<td>L cingulate</td>
<td>precuneus</td>
<td>-14 -66 32</td>
<td>4.0982</td>
</tr>
<tr>
<td>L sub-cortical</td>
<td>parahippocampus</td>
<td>-34 -46 -4</td>
<td>5.1133</td>
</tr>
<tr>
<td>R frontal</td>
<td>precentral sulcus</td>
<td>46 -6 46</td>
<td>7.7822</td>
</tr>
<tr>
<td>R frontal</td>
<td>middle frontal gyrus</td>
<td>44 12 52</td>
<td>4.0968</td>
</tr>
<tr>
<td>R parietal</td>
<td>angular gyrus</td>
<td>54 8 6</td>
<td>4.7062</td>
</tr>
<tr>
<td>R cingulate</td>
<td>anterior cingulate</td>
<td>46 -56 46</td>
<td>6.4197</td>
</tr>
<tr>
<td>R cingulate</td>
<td>intraparietal sulcus</td>
<td>52 -50 50</td>
<td>5.7924</td>
</tr>
<tr>
<td>R cingulate</td>
<td>posterior cingulate</td>
<td>4 -32 34</td>
<td>10.1873</td>
</tr>
<tr>
<td>R cingulate</td>
<td>anterior cingulate</td>
<td>6 -24 40</td>
<td>9.4472</td>
</tr>
<tr>
<td>R insula</td>
<td>anterior insula</td>
<td>36 14 -12</td>
<td>6.6248</td>
</tr>
<tr>
<td>R sub-cortical</td>
<td>putamen</td>
<td>10 16 -4</td>
<td>4.8698</td>
</tr>
</tbody>
</table>
The word rate measure had a positive effect in the bilateral anterior, superior and middle temporal lobe extending into the Temporal Parietal Junction (TPJ) in the left hemisphere. An extensive effect was also observed throughout the left middle and inferior frontal gyrus as well as antero-medial and ventro-medial left prefrontal cortex (1-4.A-D). Sub cortical regions exhibiting a positive effect included the left hippocampus and the right cerebellum. Negative effects of word rate were observed in the lateral and medial prefrontal cortex (in particular in the right hemisphere), the inferior parietal gyrus and adjacent regions, the precuneus/posterior cingulate and anterior cingulate.

1.7.2 Open class word length

Peaks of positive correlation with open class word length ($P < 0.001$ uncorrected, cluster threshold =20) are reported in table (1.3):

<table>
<thead>
<tr>
<th>Positive correlation with word length</th>
<th>Region</th>
<th>Location</th>
<th>MNI coordinates $(x,y,z)$</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L temporal</td>
<td>superior temporal gyrus</td>
<td>-58 -8 -2</td>
<td>9.0485</td>
</tr>
<tr>
<td></td>
<td></td>
<td>planum temporale</td>
<td>-66 -28 4</td>
<td>8.737</td>
</tr>
<tr>
<td></td>
<td></td>
<td>inferior temporal sulcus</td>
<td>-62 -40 12</td>
<td>7.8552</td>
</tr>
<tr>
<td></td>
<td></td>
<td>superior occipital gyrus</td>
<td>-44 -14 -22</td>
<td>5.2548</td>
</tr>
<tr>
<td></td>
<td></td>
<td>superior occipital gyrus</td>
<td>-6 -94 22</td>
<td>5.6188</td>
</tr>
<tr>
<td></td>
<td>L occipital</td>
<td>superior occipital gyrus</td>
<td>62 -6 -8</td>
<td>8.5265</td>
</tr>
<tr>
<td></td>
<td>R temporal</td>
<td>superior occipital gyrus</td>
<td>66 -26 2</td>
<td>7.8526</td>
</tr>
<tr>
<td></td>
<td>R frontal</td>
<td>fusiform</td>
<td>18 -76 -18</td>
<td>4.5062</td>
</tr>
<tr>
<td></td>
<td>R occipital</td>
<td>subcentral gyrus</td>
<td>52 -20 10</td>
<td>9.1331</td>
</tr>
<tr>
<td></td>
<td>R sub-cortical</td>
<td>anterior occipital sulcus</td>
<td>38 -66 -6</td>
<td>5.1478</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cerebellum</td>
<td>32 -64 -24</td>
<td>5.7158</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>30 -72 -24</td>
<td>5.1418</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>28 -76 -24</td>
<td>4.0613</td>
</tr>
</tbody>
</table>
Figure 1-4: Positive (red) and negative (blue) effects of word rate in the lateral (A) and medial (B) left hemisphere and lateral (C) and medial (D) right hemisphere, presented on a two dimensional representation of an 'inflated' cortex. The group t map was thresholded at \( p < 0.001 \), uncorrected, cluster threshold=20.
Word length had a positive effect in bilateral superior temporal regions (1-5.A,B) with the effect extending to middle temporal gyrus, as well as the supra-marginal gyrus in the left hemisphere. The right cerebellum also showed a positive effect of length.

Figure 1-5: Positive effects of word length in the left (A) and right (B) hemispheres presented on a two dimensional representation of an 'inflated' cortex. The group t map was thresholded at $p < 0.001$, uncorrected, cluster threshold=20.
1.8 Discussion

1.8.1 Word rate

The word rate measure produced activation in the ‘classical’ language network, including numerous regions in the temporal and frontal lobes, with the expected left lateralization of the effect in the (inferior) frontal lobe and posterior temporal (broadly Wernicke’s region) lobe. In addition positive effect was seen in the right cerebellum, which has been found to be co-active with left frontal regions in the context of linguistic tasks (Jansen et al., 2005). While traditionally the cerebellum was seen as supporting movement execution, recent imaging (Jansen et al., 2005) and patient work (Smet et al., 2007) has implicated this structure in various aspects of linguistic processing, both in production and perception. The involvement of the left hippocampus is also not surprising given its known role in semantic memory.

The negative effect (or reactivation) of word rate is also highly interpretable, implicating a network of regions known as the ‘default mode’ network (Raichle et al., 2001; Buckner et al., 2008) which includes the posterior cingulate and precuneus, the medial and lateral prefrontal cortex and posterior regions within the parietal cortex. This network is often seen to decrease its activation during cognitive tasks (and increase during rest). In all, the word rate measure has produced significant and extensive activation throughout the language network and deactivation in the expected default mode regions.

The general result is qualitatively similar to that seen by other recent experiments investigating brain response to narratives Hasson et al. (2007); Wilson et al. (2008); Awad et al. (2007). Unlike the standard approach to fMRI analysis, these results were not obtained by comparing speech to scanner noise, backward speech or other 'non-speech' conditions but instead by using a parametric modulation of the narrative itself (word rate or density). At the same time, these results share with the above mentioned literature a lack of specificity. They demonstrate that numerous brain regions are involved in the task of language comprehension, a fact that is not surprising. The problem is that they cannot tell us much regarding the specific role of the different
regions. To answer this question, we must turn to more specific dimensions of the narrative.

1.8.2 Length

One such dimension is word length. As expected, increase in word length produced increased activation in primary and secondary auditory cortices as well as the left supra-marginal gyrus which has been implicated in sublexical phonological processing (Hickok and Poeppel, 2004). In addition, activation was also observed in the right cerebellum, lending further support to the idea that it plays a role in phonological processing during perception as well. As with the word rate results, nothing is particularly surprising or new, which is in fact reassuring.

The interest of these results is that they establish that parametric modulation of extremely rapid events can be used to productively predict brain activation. For example, (Vigneau et al., 2006) have produced an extensive meta-analysis of language related imaging experiments (in the left hemisphere). The results of multiple experiments converged on 5 activation peaks associated broadly with phonology in the temporal lobe. Three of the five were specific to auditory modality while the other two were either a-modal or associated with text to phonology mapping. Interestingly all three ‘speech’ peaks fall within the activation cluster identified by the length parameter here while the two other activation peaks do not. These results, once compared to the results of the word rate measure, also demonstrate the usefulness of quantifying specific aspects of the narrative beyond the gross structure (reflected by word rate). Word length activated only a subset of the regions implicated by the general task. However, the functional implication for these regions is more specific than in the case of word rate since we have manipulated a specific property of the stimulus (length in this case).
Chapter 2

Expectation and surprisal

"If we did not have the capacity to anticipate, most of us would probably be dead" (Van Berkum et al., 2005)

2.1 Introduction

2.1.1 Pre-activation

There is little doubt that probabilistic knowledge regarding the distribution of linguistic forms plays a role in language acquisition and use. This intuition has a long history in the modern study of language and has underlay the basic methodology of structural linguistics. It was made explicit in Harris (1955), where the (type) transitional probabilities between phonemes in a corpus were used to predict morphological boundaries (cf. Saffran et al. 1996 for an experimental investigation of related ideas).

One domain where statistical knowledge regarding a linguistic form has been proposed to play a major role is pre-activation. Pre-activation is a term we chose to group together a number of psychological effects or paradigms in which past experience with a linguistic form facilitates its processing. These include the effects of priming, corpus frequency, and contextual facilitation. As the term suggests, the facilitation observed for primed, highly frequent or contextually predicted linguistic items is explained as some form of enhancement of the resting activation of the mental representation of
that unit (or components of) prior to the reception of the sensory input, compared to a lower frequency, unprimed or contextually incongruent linguistic item.

While the general explanatory framework of pre-activation has become a virtual consensus in the field\(^1\) a number of more specific questions are still debated or unresolved. In this chapter we will address some of these questions, using a novel paradigm. In particular, we will distinguish local prediction from familiarity, syntactic from lexical predictions and surprisal from uncertainty. After reviewing the relevant literature, we will present our data which is based on a novel use of a probabilistic parser to generate word by word measures for the different aspects of statistical expectation discussed here.

### 2.1.2 Familiarity and predictability

One can potentially distinguish between two types of pre-activation. The first is due to familiarity, or the number of past exposures to the relevant linguistic unit. The other is local prediction, or pre-activation due to contextual information that increases the likelihood of the relevant unit. Familiarity is usually approximated by the corpus frequency of a linguistic unit (most often a word). Local prediction has been traditionally estimated using subjective measures such as *cloze probability* (see below). More recently, computational tools have been used to approximate local prediction using conditional probability calculated from a corpus.

As we will discuss more in details in 2.2, the merit of such a distinction is under debate. One view holds that familiarity is the only primitive notion and local effects are its derivatives. The opposite view holds that prediction is the single underlying mechanism and that apparent familiarity effects are just a special case of prediction in the absence of contextual information.

In order to address the issue here, we modeled concurrently the corpus log-frequency of each word as well as its conditional probability (quantified as surprisal;\(^1\)This has not always been the case. In earlier approaches to frequency effects, for example, the explanation was not formulated in terms of activation level but in terms of order in a list, with more frequent items higher on the list and so accessed faster by a serial search mechanism.)
Hale 2001). Our results argue for a dual mechanism perspective where both familiarity and contextual predictability play a role, but in distinct (though possibly overlapping) neural networks. More specifically, our data suggests a primary role for familiarity in phonological lexical access while contextual predictability facilitates integration of the retrieved lexical item in the larger sentential structure.

2.1.3 Lexical and syntactic pre-activation

Most of the theoretical and experimental work on pre-activation has concentrated on the word and sub-lexical units. As a consequence, the role of statistical knowledge regarding syntactic structure in language processing has remained quite unexplored until very recently.

It is possible that Chomsky’s argument against the reduction of the notion of grammaticality to probability (the famous colorless green ideas sleep furiously) has been understood by researchers as a general argument against the relevance of statistical knowledge to sentence parsing (cf. Van Berkum et al. 2005). An independent issue (already highlighted by Wilson and Carroll 1954) is that while probabilistic reasoning is well defined when the set of alternatives is finite (as it is putatively the case for the lexicon), things get much more complicated (if at all possible) when the set becomes infinite (as is the case for sentences). Finally, for many years the computing power and size of available corpora where not sufficient for estimation of frequencies of units as large as sentences or even phrases.

More recently, (intertwined) developments in computing power and computational linguistics have motivated a theoretical discussion (Hale, 2001; Levy, 2008) and an empirical investigation (Demberg and Keller, 2008) directly addressing the question of syntactic pre-activation or syntactic surprisal (cf. 2.3.1). However, this growing body of work has addressed directly neither the relation between syntactic and lexical surprisal nor the relation between surprisal and syntactic complexity (cf. 2.3.3). It is possible that the behavioral measures (reading time, eye movement) used in the surprisal research up until now have limited the ability of the researchers to address these more fine grained relations.
In this chapter we report the results of the first neuro-imaging investigation of surprisal. We modeled concurrently both syntactic and lexical surprisal and observed distinct (and largely not-overlapping) brain networks activated by the two predictors. This result argues that lexical and syntactic surprisal play a role in distinct cognitive processes. In addition, we modeled syntactic complexity and found that it was hard to tease it apart from syntactic surprisal at the word level but that at the sentence level the two quantities had distinct neural ‘foot prints’.

2.1.4 Surprise and uncertainty

The final issue to be addressed in this chapter (2.9) is the effect of the word level measures discussed above at the sentence level. In particular, we will note that by averaging surprisal across a sentence we essentially produce a measure of the state of uncertainty (or entropy) induced by the sentence in the hearer.

Psycholinguistic study has traditionally concentrated on surprise or the local effect of expectation on the processing of a specific target. The related but independent effect of uncertainty or entropy has been rarely discussed or quantified. This bias away from the study of uncertainty is due in part to the use of behavioral dependent variables such as reaction time and in part to the lack of a subjective procedure to estimate the independent variable of uncertainty.

Here, the use of neuroimaging allows us to avoid the limiting factor of the behavioral measures. The use of computational tools to compute uncertainty (or averaged surprisal) overcomes the lack of a subjective measure. Our results highlight the effect of syntactic uncertainty as distinguished both from lexical uncertainty and sentence complexity.

2.1.5 Outline of the chapter

We first turn to a more detailed consideration of the issues discussed above. In 2.2 we return to the issue of the tension between familiarity and prediction based explanations of pre-activation. In 2.3 we present an informal characterization of
surprisal and discuss the distinction between lexical and syntactic surprisal. In 2.3.3 we address the intricate relation between surprisal and number of parsing steps or syntactic complexity. Finally in 2.4 we discuss the relation between surprisal and entropy and more generally the consequence of the word measures discussed here at the sentence level.

In section 2.5 we provide a brief overview of neurological correlates of expectation. The discussion focuses on electrophysiology since most relevant work on the topic has made use of Event related Potentials (ERP) measured using EEG electrodes placed on the skull. However, a short discussion of relevant fMRI literature is provided as well. Section 2.6 contains a discussion of some important issues in the design used here with particular attention to the issue of collinearity between different independent variables.

The results from the word level model are reported in 2.7 with a following discussion (2.8). The results and discussion of the sentence level measures are reported in 2.9 and 2.10 respectively. The chapter ends with a general discussion (2.11).

2.2 The source of pre-activation

2.2.1 Frequency and predictability

Frequency

The study of word processing (both in production and perception) has highlighted the relationship between a word’s frequency (or relative number of occurrences in a representative corpus of the language) and its processing (Howes and Solomon, 1951; Monsell, 1991; Dahan et al., 2001). Multiple dependent measures (e.g. reaction time, error, eye movement, ERP, BOLD) have demonstrated that high frequency words are identified or produced more rapidly and with less effort or error compared to low frequency words.

The processing of an incoming linguistic stimulus undoubtedly involves multiple stages, though researchers debate the exact nature of these stages and the degree of in-
teractions among the stages (serial versus parallel, feed forward versus feedback, etc.). Behavioral measures such as reaction time or naming latency produce a single value of the dependent variable per stimulus and so make it complicated to investigate these stages independently. During the years ingenious behavioral experimental designs (e.g. masked priming, short versus long stimulus onset asynchrony priming) as well as more fine grained dependent measures (eye tracking, ERP, MEG) have permitted to appreciate the contribution of different linguistic variables to temporally distinct stages of processing.

One finding which seems to be supported by multiple evidence (cf. Pylkkänen et al. 2002; Hauk and Pulvermüller 2004 for discussion) is that the frequency of a lexical item correlates (negatively) with dependent variables tapping both early (200 – 350 ms) and somewhat later (> 400 ms) stages of the processing of incoming linguistic stimulus (usually a word). It is logically possible (and indeed assumed by a number of lexical access models) that these early and late effects arise for different reasons.

The early relationship has most often been interpreted as an effect of familiarity with the specific lexical item (or sublexical properties of that item). That is, frequency approximates the familiarity of the system with the specific lexical item. More familiar items have a higher resting activation and so are accessed more easily or faster.

The late effect of frequency has been often interpreted differently. Usually, this effect is considered to arise due to post lexical access competition between the retrieved lexical item and strongly activated competitors. The frequency of a lexical item appears to be a good approximation of the extent of that late competition. Low frequency lexical items are subject to stronger competition from the alternatives and so are harder and take longer to process.²

²We are purposefully being vague with respect to the identity of the competitors and the nature of the competition. Models differ with respect to these issues in ways that we are not able to address in the current work. Since the empirical generalization relating frequency to post lexical competition is common to the different proposals, we will keep our discussion here at this rather descriptive level. We hope to address the more specific details of this late competition in future work.
Contextual prediction

A different line of research, which historically stemmed from the attempt to define measures for text comprehension difficulty (often for educational or testing purposes), has focused on the effect of the predictability of a word in a given context on its processing. Taylor (1953) developed a procedure to calculate the cloze probability\(^3\) of a word in a context. A group of subjects read a text in which every nth word was replaced by a line and were asked to fill in the missing words. The probability of a word to be suggested in a specific context was termed its cloze probability. Taylor himself demonstrated that a text's difficulty or comprehensibility is predicted by the (averaged) cloze probability of (a sample of) the words it contains. Later research has demonstrated that cloze probability predicts reading time (Stevens and Rumelhart, 1975), eye movement patterns (Rayner and Well, 1996) and brain response amplitude (Kutas and Hillyard, 1984).

It is possible to logically distinguish two consequences of prediction (paralleling the early versus late effect of frequency). The first is the effect prediction can potentially have on the pre-activation of the target item which could facilitate its access (similarly to familiarity). A second effect could be that of competition. Prediction can pre-activate a 'wrong' lexical item, hence increasing post lexical access competition. This distinction has not received much attention, but recent work by Federmeier et al. (2007), to be discussed below, provides ERP evidence supporting it.

Despite the conceptual similarity of frequency and predictability, these two lines of research have been mostly pursued separately. This was probably a consequence of the different domains of inquiry (single words versus whole sentences or even larger texts) and the different estimation procedures (frequency was calculated using a corpus while predictability was calculated using a subjective measure). Reviewing the literature, it is possible to identify two perspectives regarding the relationship between frequency and predictability. In some cases, this question has been the explicit topic of inquiry (Van Berkum et al., 2005) but in others it has been implicit in the assumptions of

\(^{3}\)the term cloze was coined as a derivative of the gestalt term closure, or the tendency of an observer to complete partial patterns.
The top-down view: it is all about prediction

Under the ecological perspective proposed by Anderson and Schooler (1991), the reason human memory exhibits word frequency effects is tied to the fact that "the probability that a memory will be needed also shows reliable relationships to frequency..." (p. 396). In other words, this perspective highlights the importance of statistical distribution of past events as a natural source for predictions regarding the future. Applied to language comprehension, the frequency effect could be re-conceptualized as a special case of predictability in the absence of any contextual biases (e.g. when a word is presented in isolation, as in most psycholinguistic paradigms). Priming (under this view) is an effect of prediction generated by a very minimal context (a single word, the prime).

According to this view, frequency effects should be reduced or even eliminated when a word is processed in a constraining context, such as a sentence. This prediction has received corroboration from a study by Van Petten and Kutas (1990). Using ERP, these authors demonstrated that the N400 component (a negative deflection of the ERP signal occurring around 400 ms after the onset of word presentation) is modulated by word frequency only when the word is sentence initial (with no context). The N400 response to sentence medial words was not modulated by frequency.

As we will discuss below, it is important to be careful when interpreting this finding (and other null effects in ERP). The fact that an ERP measure cannot detect an effect of frequency in sentence medial words may reflect the underlying neural reality (and in that case support the prediction perspective). However, it is also possible that either because the rather late onset of the N400 or because it most probably reflects activation in multiple cortical loci, the N400 is not sensitive enough to detect the relevant frequency effect (cf. 2.5.1).
The bottom-up view: familiarity without prediction

Other theoretical models, in particular models of lexical access (Forster, 1979; Marslen-Wilson, 1987; Norris, 1994), have espoused (either explicitly or implicitly) the opposite view, according to which prediction is not an explicit goal of the model (often described as bottom-up view). Familiarity (approximated by frequency) is a central component in these models since it determines the resting activation of each lexical item in long term memory, which in turn determines how fast it is accessed or how dominant it is in a competitive view of lexical access (Marslen-Wilson, 1987).

Priming, under this view, is a recency effect that temporarily boosts the resting activation of the lexical item or some sub-structure (morphological, phonological or semantic) of that item. There is no computational goal for priming, it is simply a side effect of the memory retrieval algorithm.

Evidence for this position comes from the processing of ambiguous words in context. Swinney (1979), using cross modal priming, demonstrated that one of the meanings of an ambiguous word can still prime a semantically related target even if the word is presented in a sentence which clearly favors the other meaning. Crucially, this effect held only when the visual target was presented simultaneously with the ambiguous auditory word. Other experiments, both those using behavioral measures as well as ERP, have replicated and nuanced this original finding. Directly relevant to the discussion here is a finding by Tabossi (1988) that priming by the contextually unsupported meaning can take place only if this meaning is more frequent (or dominant). This result strongly argues for an early effect of frequency which is unaffected by predictability (cf. Sereno 1995).

It is important to note (and this issue bears directly on the results here) that bottom-up and top-down interpretations of frequency differ with respect to their explanation of of the early effect of frequency (familiarity versus prediction). That is, the effect of frequency on the perception of the linguistic stimulus and the recovery of the identity of the lexical item associated with that input (lexical access). The

---

4In particular short SOA priming. We will return to this issue below.
interpretation of the late effect is not necessarily subject to the same dichotomy. Since there is no debate that executive control or contextual information plays a role in later stages of linguistic processing (integration of the retrieved lexical item into the larger sentential structure) both top-down and bottom-up perspectives are compatible with effects of prediction on post-lexical access competition.

2.2.2 Distinguishing prediction and familiarity

The first goal of this chapter was to quantify the effects both of prediction and familiarity using fMRI (for the first time, to our knowledge) and to attempt to distinguish the top-down from the bottom-up perspective on the frequency effect. This was done by modeling concurrently both the local predictability of each word (its conditional probability) and its corpus frequency.

Brain activation explained uniquely by conditional probability can be interpreted as related to top-down prediction. Brain activation explained uniquely by frequency is more problematic to interpret given the lack of fine grained temporal resolution in fMRI. A unique frequency effect could be evidence for an early familiarity effect but could also index later competition.\(^5\)

In order to attempt to resolve this ambiguity we can use prior knowledge regarding the region(s) exhibiting the effect. A second source of information is the Sentence model where we modeled both averaged lexical surprisal and averaged frequency. It is possible to use the following reasoning: Brain activity that exhibits frequency effects at the word level and the sentence level is modulated by competition. This is the case since competition is induced by explicit prediction and so when averaged it reflects uncertainty (or entropy; 2.4). Conversely, brain activity that exhibits frequency effects at the word level but not at the sentence level is sensitive to familiarity and not competition. This is the case since familiarity is conceived of as a bottom-up property and does not reflect expectations or other aspects of executive control and so when averaged, is not expected to produce a global measure of uncertainty.

\(^5\)MEG data could be used eventually to directly distinguish early versus late effects.
2.3 Syntactic versus lexical prediction

Most work cited above focused on the prediction of the specific lexical identity of the next word (most often a noun) given the sentential or discourse context. However, less is known about the extent of prediction of the upcoming syntactic structure. This question regarding what is being predicted can also be framed as a question of scope. The kind of (lexical) prediction usually put to test can be thought of as local. Is this local prediction the goal of the computation or a side effect of a less local (syntactic/sentential) prediction regarding a larger structure, such as the ultimate form of the sentence?

From the perspective of speech recognition, predicting the upcoming word is very useful in order to improve detection and recognition given the noisy channel. Automatic speech recognizers, for example, rely heavily on such predictions, usually derived from Hidden Markov or n-gram models (Rabiner, 1989). Neural network modeling of sentence processing (Stevens and Rumelhart, 1975; Elman, 1990) takes the quality of local word prediction as the success measure.

However, from a communicative, or message level perspective (as made explicit in Information Theory and in theoretical linguistics) individual words form part of a larger structure (the sentence or the message) and it is through this role that words inherit their predictability (and more generally, function). Though this view has been present in the theoretical discussion as well as in Natural Language Processing, it has had less of an effect on the experimental psycholinguistic literature (Van Berkum et al., 2005; Levy, 2008).

A possible reason for why syntactic prediction as distinct from lexical prediction has been rarely investigated has already been discussed in an early paper about the role of Information Theory in psycholinguistics (Wilson and Carroll, 1954). While lexical predictions can be estimated by subjective ratings, an equivalent measure for syntactic predictions is not obvious. The alternative, computational, method to compute predictability requires both a very large corpus and a probabilistic parsing model that can generate, for each word, all the possible syntactic continuations and
rank them by probability. It wasn't until the 1990's that such computational tools were available and not until 2000 that they were put to use for this purpose. The ability today to estimate predictability from a corpus instead of (or in addition to) relying on subjective judgments has allowed for the computation of syntactic expectations.

The second goal of this chapter is to estimate the relative contribution of lexical (or local) and syntactic (or global) predictions to on-line processing. In the case that we find that both play a role, we would like to use brain activation distribution to determine whether the two types of predictions are supported by the same neural network or separate ones. We turn now to the discussion of the computational measure used here, namely surprisal.

### 2.3.1 Surprisal

In an influential paper, Hale (2001) proposed a psycholinguistic framework founded on the view of prediction as a global, message level, task. Building on the development of incremental probabilistic parsers for Natural Language Processing (Stolcke, 1995), Hale proposed to derive the conditional probability of a word given the context using a probabilistic parser. The (minus) log of this probability was named *surprisal*. This quantity, in effect, measures the change in probability mass of the parse due to the disconfirmation of structural predictions when a new word is processed. In other words, the surprisal level of a word corresponds to the probability mass of analyses (or parses) which were consistent with the preceding string before the addition of the word and which are now inconsistent.

A probabilistic incremental parser generates (or can be made to generate) a probability for every prefix (or partial string). If the parser is parallel, the probability of the prefix could be viewed as the sum of the probabilities of all possible parse trees that could generate that prefix (or alternatively all the analyses in the beam\(^6\)).

In general, when the prefix is extended, its probability reduces since the number

---

\(^6\)Since every prefix is compatible with potentially infinite partial derivations, parsers often maintain only a subset of these possibilities, known as a beam.
of potential parses compatible with it reduces. For Hale, the surprisal value of a word is equal to the extent to which it has reduced the probability of the prefix (measured as the ratio between the probability of the prefix at word \( w \) and the probability of the prefix at word \( w-1 \)). Surprisal shifts the target of expectation from the word itself to the identity of the final sentence. A word is more or less surprising to the extent that it informs the processor regarding the eventual outcome of the sentential parse.

\textbf{2.3.2 Syntactic versus lexical surprisal}

From the perspective of incremental probabilistic parsing, an input word makes potentially two distinct contributions to the partial derivation. At least in the case of open class words, the input word contributes the specific meaning/concept associated with it in the mental lexicon. However, the input word also provides information regarding the syntactic structure. Take for example the two alternative continuations in 5:

\begin{enumerate}
\item John knew Mary ...
\item a. well
\item b. left
\end{enumerate}

The two continuations differ quite obviously in the lexical meaning associated with them, but also with respect to the constraints they impose on the global sentential structure. \textit{Well} supports an interpretation of \textit{Mary} as the direct object of the verb \textit{know}. \textit{Left}, in turn, supports an embedded subject interpretation of \textit{Mary}. Importantly, the two dimensions are logically independent.

In the cloze-probability literature, experiments keep the sentential context and word class of the target item identical across conditions, varying only the specific lexical item. Reinterpreting this literature within a surprisal framework, we could describe these experiments as manipulating \textit{lexical} surprisal. The conditional probability of the specific lexical item given the preceding context \textit{and} the identity of the non-terminal dominating that lexical item (the part of speech tag).
Most work explicitly grounded within the surprisal perspective (Hale, 2001; Narayanan and Jurafsky, 2002; Hale, 2006; Levy, 2008; Demberg and Keller, 2008; Vasishth et al., 2008) has been interested in syntactic surprisal and has made use of un-lexicalized probabilistic context free grammars (PCFGs) to predict surprisal values. This line of research has attempted to model structural prediction (of rules or part of speech tags) independently of the specific lexical items instantiating these structures. Recently, Hale and colleagues (Vasishth et al., 2008) demonstrated that syntactic surprisal and empirical predictability (a measure similar to cloze probability) are not correlated. These authors did not compute lexical surprisal but it is plausible that lexical surprisal would have been more correlated with empirical predictability.

While the effect of lexical prediction (or lexical surprisal) is by now well established (Kutas and Federmeier, 2000; Van Berkum et al., 2005), a distinct unambiguous effect of syntactic surprisal has not yet received strong confirmation in the experimental literature. Demberg and Keller (2008), using a linear mixed effects model to analyze a naturalistic eye-tracking corpus, found an effect for syntactic surprisal (coefficient=1.03), however the combined coefficient of non-syntactic surprisal\(^7\) was one order of magnitude larger (10.17).

Levy (2008) reanalyzes a number of published results in an attempt to motivate a syntactic surprisal account. The problem is that at least some of these cases are not unambiguously syntactic but might have a lexical nature. Consider the results reported by Konieczny and Doring (2003). These authors trained a Simple Recurrent Network (SRN) on a subset of German. The network was then exposed to novel, verb final, sentences in which the number of preverbal arguments was varied. The network’s prediction of the sentence final verb improved with the increase (addition of a dative NP) in the number of arguments. The same manipulation was used in sentences in an eye tracking reading experiment with German native speakers:

\(^7\)The authors did not model lexical surprisal but included a number of regressors that arguably reflect this property, namely bigram transitional probability and frequency (cf. Frisson et al. 2005). Since these two measures only approximate lexical surprisal, an explicit modeling of lexical surprisal might have had an even stronger effect.
Die Einsicht, dass der Freund dem Kunden dat das Auto aus Plastik verkauft, erheiterte die anderen.

"The insight that the friend sold the car made from plastic to the client amused the others."

Die Einsicht, dass der Freund des Kunden gen das Auto aus Plastik verkauft, erheiterte die anderen.

"The insight that the friend of the client sold the car made from plastic amused the others."

Die Einsicht, dass der Freund dem Kunden dat das Auto aus Freude verkauft erheiterte die anderen.

"The insight that the friend sold the car to the client just for fun amused the others."

In accordance with the predictions of the SRN, the addition of a dative argument (compare the dative *dem Kunden* in 6 with the genitive *des Kunden* in 7) reduced reading time of the verb (*verkauft*). This result could be seen as a syntactic surprisal effect (as proposed by Levy) since before seeing a dative argument, the probability mass regarding the next word (or part of speech) is distributed between (minimally) an additional NP (dative) and the verb. However, once a dative has been incorporated into the structure the next word cannot be a dative and so the probability of a verb increases.

However, there is an alternative, lexical, explanation (the one put forward by the authors themselves), namely that the addition of a dative constrains the possible class of verbs and so facilitates lexical access for the specific verb. In this respect it is useful to take into account the results of the second manipulation reported by Konieczny and Doring (2003). In their eye-tracking experiment a modifier phrase preceded the clause final verb in all sentences. The authors compared regression path duration (first pass reading) of the verb with a preverbal modifier (*aus Freude* in 8)
to sentences where the modifier modified the preceding noun phrase (*aus plastik* in 6).

Though Levy does not discuss these results, it is possible to extend the logic of the surprisal predictions to this contrast. A verb is less surprising following a verbal modifier than a nominal modifier since the verbal modifier indicates that the preceding noun phrase has been completed (while a nominal modifier does not) and moreover, two verbal modifiers are rare and so after one verbal modifier the parser would not expect a second one. Lexical surprisal on the other hand does not predict a difference unless the preverbal modifier specifically restricts the choice of the following verb. Crucially, the authors found no difference in eye movement across modifier type (8 versus 6). This null result, taken together with the significant difference due to the presence of a dative suggests that lexical surprisal rather than syntactic surprisal is the driving force in these cases of head final integration.

Summing up, lexical and syntactic surprisal are (potentially) conceptually distinct, related to distinct domains of prediction. Empirically the two have been investigated

---

8Levy also reanalyzes data from a similar experiment in English (Jaeger et al., 2005) which is probably the clearest case of syntactic surprisal to date. These experimenters manipulated the number of post nominal modifiers of the subject NP. As expected from a syntactic surprisal perspective (and confirmed by Levy using a PCFG), increasing the number of post nominal modifiers decreased reading time on the verb (since by increasing the number of modifiers, the likelihood of the next word to be a verb increases). Interestingly, the reading time decrease was not linear with the number of modifiers, while the surprisal predictions were. In any case, it is hard to generalize from one very specific syntactic construction regarding the general role of syntactic surprisal.

9Levy (2008) argues that an important advantage of surprisal over other measures of parsing complexity is its robustness to the choice of generative mechanism (for example HMM versus PCFG), and representational assumptions. "Any two generative stochastic processes that determine the same set of conditional word probabilities will make exactly the same predictions about processing difficulty, regardless of the representational content of these processes or even the nature of the underlying (hidden) structures within the process" (page 7). From the discussion above regarding the lexical/syntactic distinctions regarding surprisal there seems to be a problem with Levy's perspective. An HMM and a PCFG trained on the same corpus might in principle produce the same conditional probability for a particular word w given a prefix p. However, the PCFG predictions can be broken down into the syntactic and lexical surprisal of w given p while in an HMM the two
separately. Lexical surprisal (under its different guises) has been repeatedly shown to play an important role in processing. Whether syntactic surprisal plays a role as well is still largely an open question.

In the experiment here we quantify both lexical and syntactic surprisal. Syntactic surprisal was calculated as the conditional probability of the part of speech given the preceding context. Lexical surprisal was calculated as the conditional probability of the particular lexical item given the part of speech and the preceding context.

2.3.3 Surprisal and syntactic complexity

An important issue regarding syntactic surprisal is the delicate relation in a PCFG (as well as other probabilistic grammars) between syntactic complexity and probability. In a PCFG, each rewrite rule is assigned a conditional probability (the probability of the right hand side given the left hand side context, possibly with additional information such as the lexical head as in Collins 1997 or the identity of the dominating category or c-commanding lexical heads as in Roark 2001). The probability of larger tree fragments is the product of the rules used to generate them. As a consequence, each additional rule (tree extension) will automatically reduce the probability of the structure. In other words, there is a potential confound between derivation length (per word) and the syntactic surprisal value associated with that word.

Under a simplistic (and quite natural) assumption that derivation steps are computations and that computations entail cognitive cost, an apparent syntactic surprisal effect might actually be due to (at least in part) computational complexity\(^\text{10}\). This issue has not been addressed in the surprisal literature and it is an open question how to tease apart surprise from complexity (or even if such separation is possible).

\(^{10}\)It is of course possible that the human parser chunks sequences of rules with high conditional probabilities (for example, the rule that expands VP to V NP with the rule that expands NP to D N, cf. Johnson et al. 2007 for one such formalism). If this is the case than there would not be a necessary relation between number of rules and actual computational steps (Michael Collins, p.c.).
Levy (2008) makes the point that surprisal does not require us to commit to a particular generative formalism or other assumptions regarding the exact internal representations that generate the parser's predictions. This perspective is obviously relevant here. Complexity measures will vary from one grammar to another and will depend on parser specific implementation. It is most probably not the case that the Penn tree bank categories used by Roark's parser correspond to the 'true' constituents manipulated by our mental grammar. Moreover, as has been argued in Chomsky (1957) (cf. also Shieber 1985), context free grammars are not sufficient to capture the complexity of natural language. At the same time, it is possible that a parser specific complexity measure might correspond to parser independent grammatical facts. It seems to us that at the end of the day the appropriateness of, and causal relation among, these different measures is an empirical question. We chose to model syntactic surprisal with concurrent modeling of parser steps per word in order to get a sense of the relation between the two quantities as captured by brain activation.

### 2.4 Surprise and entropy

Work looking at the cognitive effects of predictions and expectations has concentrated on the effect of surprise, or the extent of compatibility of an input with prior expectations. This is true for the study of lexical expectations, which are most often measured indirectly via their effect on the processing of a target item such as the N400 effect (discussed more below). It is also the case for the study of syntactic surprisal, as in the study by Konieczny and Doring (2003) discussed above. Interestingly, this body of work did not address the related but distinct question of entropy or uncertainty.

The surprisal perspective holds that human incremental parsing relies on the formation of predictions regarding downstream input and the overall structure of the sentence. The existing literature has mostly explored the nature of this predictive

---

11This holds also of the study of the mismatch negativity effect, the P300 and other manipulations of predictability in non-linguistic contexts.
mechanism by manipulating the predictability of the target item. The basic question has been the correlation between a dependent variable (whether behavioral or neural), measured at the target, and the predictability of that target. However, one could also investigate this mechanism by correlating the dependent variable with uncertainty.

Hale (2006) replaces the notion of surprisal with entropy reduction (proposed by Wilson and Carroll 1954). Under this view, the cognitive load associated with a word is not how surprising it is (as in Hale 2001) but rather to what extent it reduces our uncertainty regarding the final output of the parse (or the end of the sentence as the title of his paper puts it)\(^{12}\). Notice, however, that though Hale uses the uncertainty state of the parser to calculate this quantity, the state of uncertainty itself is not modeled.

Alternatively, we could ask, at every stage of the incremental parsing process, how certain is the parser regarding the identity of the next item or the eventual output of the current, partial, parse\(^{13}\). A priori, it is possible that this uncertainty state itself will have neuro-cognitive correlates. For example, increase in uncertainty might lead to increase in top-down attention modulation.

The question whether entropy has a neuro-cognitive impact during sentence comprehension is also directly relevant to the top-down/bottom-up debate addressed earlier. In a parsing architecture that does not actively engage in prediction there is no a-priori expectation to find any entropy effect since (un)certainty is not explicitly modeled in the system. On the other hand, entropy is an integral notion in a cognitive architecture that makes explicit use of prediction.

In effect, recent fMRI experiments have started to explore the effect of entropy in non-linguistic paradigms (Strange et al., 2005; Overath et al., 2007; Volz et al., 2005). Strange and colleagues used a mixed blocked/event related design. Subjects viewed

---

\(^{12}\) Though similar and related to surprisal, entropy reduction can make, in principle, different predictions since a highly surprising event can increase entropy, not necessarily reduce it.

\(^{13}\) Work looking at lexical or structural ambiguity has touched upon the topic of entropy but not over the end of the sentence. Entropy related to ambiguity measures the (un)certainty of the parser regarding the correct analysis of the already accumulated input string. We will return to this literature in chapter ??.
sequences of shapes drawn from a distribution that varied across blocks. The experimenters used an ideal Bayesian observer to assign a surprise value for each event (depending on the block specific probability of drawing that shape). The blocks differed from each other in entropy, or the overall certainty the ideal observer had regarding the correct distribution of shapes in that block (a running average of surprise). This clever design allowed the experimenters to distinguish between the effects of surprise (at the event level) and entropy (at the block level). Strange and colleagues found that surprise was positively correlated with activation throughout the visual processing stream while entropy correlated with activation in the left hippocampus.

Mathematically, entropy is calculated as an average of the surprisal (or conditional probabilities) of a set of events weighted by the probability of these events (Wilson and Carroll, 1954). The set of events can be defined paradigmatically or syntagmatically. If we are interested in the entropy state at a specific point regarding the next step, the relevant, paradigmatic, set of events would be all possible next immediate steps. If, instead, we are interested in the uncertainty associated with a larger time interval, the relevant syntagmatic set of events will contain the events that took place during that time interval.

Strange et al. (2005) computed entropy syntagmatically, averaging the surprise value of all the events that occurred during the block. Here, we quantified entropy in a similar way, averaging syntactic and lexical surprisal as well as word frequency per sentence\textsuperscript{14}. These entropy measures were included in the Sentence model.

\subsection{2.4.1 Syntactic entropy and complexity}

In the discussion above we discussed the relation between syntactic surprisal and syntactic complexity (or number of parser steps). Since the measures are correlated, we entered both into our Word model.

The same holds at the sentence level. Averaged number of parser steps can be conceived of as a metric of the syntactic complexity of the sentence. A higher number

\textsuperscript{14}The paradigmatic entropy measure will be incorporated into our work in the near future.
of parser steps per word indicates a more articulated syntax for the same string length. This measure is very similar to the measure of syntactic complexity proposed by Yngve (1960). In parallel to the Word model, we added averaged number of parser steps (or the syntactic complexity of the sentence) to the Sentence model.

2.5 Brain correlates of expectation during sentence processing

In this section we briefly review relevant findings from neurolinguistics to the question of pre-activation. We begin with an overview of two ERP components (The N400 and ELAN) and then turn to relevant findings from fMRI.

2.5.1 The N400

The N400 is a negative deflection in the ERP signal centered at around 400 ms after the onset of word presentation. Kutas and Hillyard (1980) originally identified this component using a sentential violation of expectation paradigm. Sentence final incongruent words elicited a larger N400 negativity compared to congruent words. Since 1980, multiple experiments have continued to explore different aspects of this component in sentential, as well as non-sentential, contexts. In what follows we will not attempt to provide a review of the literature (cf. Van Petten and Luka 2006; Lau et al. submitted for recent reviews and synthesis). Instead, we will touch on specific issues that are of relevance to the experiment here.

The N400 was found to be modulated in a variety of pre-activation paradigms. (Kutas and Hillyard (1984)) manipulated cloze probability of the last word in a sentence and found that words with higher cloze probability produced a reduced N400 component. Van Petten and Kutas (1990) found that word frequency modulated N400 in sentence initial context (only). The N400 response to high frequency words was attenuated. Finally, (Holcomb and Anderson, 1993) demonstrated that semantic priming also attenuates the N400. It is important to note that in all these cases the
difference across conditions reflects a *facilitatory* effect of the 'pre-activated' condition rather than some form of error detection for the 'non-pre-activated' condition. This is the case since the N400 profile for the latter is identical to the N400 for words with no context.

Is the N400 modulated by (bottom-up) familiarity or by explicit prediction (cf. 2.2)? This is not an easy question to answer. Van Petten and Kutas (1990) found that the frequency effect on the N400 does not show up for sentence medial or final words. This finding suggests that frequency and local predictability are inter-related (but see below). However, it does not distinguish between a bottom-up contextual effect and an explicit prediction. St. George et al. (1994) have demonstrated that N400 is modulated by global inferences from the context and not through low level priming from related words in the context. However, while this finding argues against a word-to-word priming, it does not rule out a more sophisticated (and indeed plausible) account of priming where a complex linguistic representation (or context) can prime a lexical item.

More recently, DeLong et al. (2005) have claimed that the N400 response to the indefinite determiner (a versus an) is modulated by the cloze probability of the following noun:

(9) The day was breezy so the boy went outside to fly

a. *a* kite
b. *an* airplane

Airplane is a less expected completion of 9 as established using the standard cloze measure. Importantly, DeLong et al. (2005) found an increased negativity already at the determiner *an*, suggesting that not only the lexical noun itself has been subject to anticipation but rather the entire noun phrase. A bottom-up priming account would be hard pressed to account for such data since there is nothing in the context itself to pre-activate preferentially *a* over *an*. In order to account for the difference observed between the two determiners, one must appeal to the phonological shape of the noun (though it has not been actually present in the input).
While these results argue for explicit lexical prediction, the N400 effect that was observed on the determiner was earlier, of shorter duration and of different scalp distribution compared to the N400 observed for the noun. In addition, Van Berkum et al. (2005), who performed a similar manipulation (using gender agreement on a pre-nominal modifier in Dutch), did not find an increased negativity for the less predicted agreement but rather an early positivity. The two experiments differed on a number of dimensions (modality, language, type of cue, etc.) however the ambiguous status of De long et al’s effect combined with the lack of an N400 effect in Van Berkum et al’s study suggests that prediction, though real, is not indexed by the N400.

An interesting feature of the DeLong et al. (2005) study was the use of a parametric design with 10 levels of cloze probability. The authors demonstrated high correlation between the ERP measure at the N400 window (300-500) over centro-parietal regions and cloze probability. The gradient modulation of the N400 by cloze probability argues against another interpretation of the N400 as a post lexical access competition measure. In a recent elegant experiment Federmeier et al. (2007) provide further and more direct evidence for this view. The authors measured ERP response in a two by two design:

(10) He bought her a pearl necklace for her..
   a. birthday
   b. collection

(11) He looked worried because he might have broken his...
   a. arm
   b. collection

They compared ERP response to high versus low cloze probability words in strongly constraining contexts (10a versus 10b) where most of the cloze probability mass is absorbed by a single word (birthday) and weakly constraining contexts (11a versus 11b), where there is no one highly preferred completion. Low cloze probability items generated a stronger N400 response compared to the high cloze probability independently of the context type.
These results confirm that the cognitive process underlying the N400 component is facilitated by predictive information in the context but is not sensitive to the presence of competing predictions. The emerging picture is that the N400 reflects lexical access which is influenced by top-down expectations (that might begin before the word is even presented) but not post lexical integration costs at least to the extent that these involve competition.

Since the N400 component is manipulated by a variety of pre-activation paradigms and, moreover, exhibits an interaction effect between frequency and context (Van Petten and Kutas, 1990), can we conclude that a single top-down pre-activation mechanism is at play? This conclusion is not warranted for at least two reasons.

First, the N400 appears to have multiple cortical sources. The standard low electrode density (< 40 electrodes) setting of ERP is not ideally suited for localization purposes (cf. Halgren et al. 2002 for discussion). Alternative localization methods (MEG, Halgren et al. 2002, fMRI, Friederici et al. 2003, intracranial recordings, Nobre and McCarthy 1995) have produced a rather inconsistent body of results (cf. Lau et al. submitted; Van Petten and Luka 2006).

There seems to be general agreement that regions within the (anterior) medial and superior left temporal lobe participate in the generation of the N400. The role of the inferior frontal cortex (IFC) is less clear. MEG and ERP localization using Equivalent Dipole models usually don’t identify sources in that region but this might well be due to the limits of the method. At least two studies using distributed source analysis (Halgren et al., 2002; Maess et al., 2006) have identified increased activation for either anomalous or less expected words in the IFC within the N400 time window. Halgren et al. (2002), contrasting incongruous and congruous sentence final words, found a posterior-to-anterior spreading activation (or difference) wave. Starting at the planum temporale around 250 ms, activation spreads to the superior and anterior temporal lobe by 270 ms followed by the orbital (or ventral) part of the inferior frontal cortex. Activation spreads to the dorsolateral inferior frontal (Broca's area) by 370 ms.

It appears, then, that the N400 component is in fact a complex summation of
neural activation across a number of distinct brain regions. This conclusion argues against a 'one-component equals one cognitive process' view. It is then possible, that the N400 umbrella covers a number of distinct cognitive processes where some could potentially be modulated by bottom-up familiarity and some by explicit prediction.

Second, the N400 could reflect, a relatively late effect which is affected by a number of, possibly disjoint, earlier effects. Hauk and Pulvermüller (2004), for example, found a frequency effect before 200 ms. An equivalently early effect of constraining context has not been demonstrated. This finding is in line with eye tracking results that have demonstrated an early effect of frequency which was not affected by contextual predictability.

Summing up, the N400 appears to be sensitive to the pre-activation level of a lexical item. It does not index competition or error detection. While the fact that a variety of sources of pre-activation influence the N400 appears to argue for a unified account of pre-activation, the spatial heterogeneity and the relatively late onset of the N400 do not permit this inference.

2.5.2 N400 and the late positivity

Throughout the years, a number of papers investigating the N400 effect have reported a later, usually frontal, positivity in the incongruent or less predictable conditions. This effect has been often noted but not integrated into the cognitive architecture under discussion until very recently. Van Petten and Luka (2006); Lau et al. (submitted) propose that this later effect indexes integration with the preceding context and in particular competition with top-down (or more strongly) predicted alternatives.

To this effect Federmeier et al. (2007) found that though the N400 response to low cloze probability items was not affected by the presence of absence of a competitor (10b versus 11b), the latter condition (10b) did elicit a later positive shift over left anterior electrodes. This finding is consistent with the suggestion, discussed earlier, that competition has a late effect (post N400) on processing. Van Petten and Luka (2006) suggest that the late (370 ms) MEG activation in the dorsal aspect of the inferior frontal region observed by Halgren et al. (2002) might reflect this later stage
of processing. In contrast, the earlier activation observed at the orbital region would be part of the N400 proper, reflecting lexical access.

The association of the dorsal aspect of the inferior frontal regions with mediation of competition, and the orbital with semantic retrieval is in line with a recent model based on data from non-linguistic tasks (Badre et al., 2005). The data regarding this later component (and its relation to other known component such as the Left Anterior Negativity, P300, or the P600) is still too sparse to draw any firm conclusions.

2.5.3 The Early Left Anterior Negativity (ELAN)

Neville et al. (1991) and Friederici et al. (1996) described an early left anterior ERP component (around 200 ms) that was sensitive to grammatical category violations as in (12a,13a):

(12) a. * The scientist criticized Maxs of proof the theorem.
    b. The scientist criticized Maxs proof of the theorem.

(13) a. * Die kuh wurde im gefüttert.
    the cow was in-the fed
    the cow was in-the barn fed
    (from Lau et al. 2006)

Friederici and others have argued on the basis of this finding that (bottom-up) basic structure building operations (basic phrase structure rules) take place extremely fast.

Lau et al. (2006) have advanced a different view according to which the ELAN is not a consequence of a bottom-up analysis of the input followed by error detection, but rather a result of disconfirmation of prior expectations or predictions. These authors demonstrated that manipulation of the strength of predictions in sentences like the one in (12a) has an effect on the strength of the ELAN response to the same word order/grammatical class violation. This finding suggests that the ELAN might be the syntactic analog of the N400, exhibiting sensitivity to the degree of predictability of the syntactic category or structure. In effect, Friederici’s group often used a determiner followed by a participial (instead of a noun) to generate an ELAN
response. This word order is not ungrammatical in German (as long as the noun appears later) but is certainly less common, strengthening the view of the ELAN as a prediction based effect. To our knowledge this interpretation of the ELAN has not yet been explored to a greater depth in the literature. One question that could be asked (along the equivalent distinction made regarding lexical prediction) is whether the ELAN reflects the processing of unexpected syntactic continuation or the competition between the actual continuation and the predicted one.

With respect to localization, Friederici et al. (2000) argue, based on fMRI constrained MEG data, that the ELAN's source is to be found in the superior temporal lobe and the (left) temporal pole, the most anterior portion of the temporal lobe. Recent work in Pylkkänen's lab (Dikker et al., Accepted) explored MEG correlates of grammatical violation of a similar kind to (12a) but in the visual modality. These authors found an ELAN like response (temporally) in the visual cortex. This finding led them to suggest that the ELAN is a modality specific response, due to top-down modulation of specific (auditory/visual) form expectations.

2.5.4 FMRI

Sentential manipulations

FMRI investigation of predictability and surprise in the course of sentence processing has most often attempted to replicate the N400 paradigm by manipulating sentence final congruency (cf. Friederici et al. 2003; Kuperberg et al. 2003 among others). Lau et al. (submitted) conducted a meta-analysis of 17 such studies (looking only at left hemisphere activation). Most studies found an effect of in-congruency in the temporal lobe but there was little overlap with regard to the specific sub-region implicated across studies. All but two of the studies found a congruency effect in the IFC. BA 45, or the pars triangularis, was implicated by the majority of the studies (n=10). Activation in the other two sub-regions (BA 44 and 47) was less consistent across studies.

While these findings seem to stand in contradiction to the electrophysiological
evidence (which localize the N400 primarily to the temporal lobe) it is important to keep in mind the nature of the BOLD signal. The studies reported in the meta-analysis did not attempt to distinguish early and late neural effects of in-congruency. As a result, the effect seen in BA45 could reflect post lexical access competition (as was suggested for the late positivity discussed above). The lack of consistent localization in the temporal lobe could be due to a number of reasons. One prominent speculation (Richard Wise, CNS 2008 address, Spitsyna et al. 2006) is that the problem lies in the sensitivity of the canonical EPI acquisition sequence, used in most fMRI experiments, to artifacts in the ventral medial temporal lobe, which is consistently implicated in semantic tasks by intracranial recordings and PET.

Another possible explanation could be the effect of subsequent (> 400ms) differences across conditions on overall BOLD effect. The electrophysiological paradigm allows the researchers to ignore differences across conditions downstream from the 400 ms region. This privilege does not extend to the fMRI paradigm. Semantic incongruency may cause decrease in neural activation compared to control at a later point during processing, possibly due to failure of composition (cf. Weiss and Mueller 2003 for suggestive data). As a result the overall BOLD effect might not increase in the implicated temporal regions.

**Semantic priming**

FMRI investigation of single word priming have revealed an interesting pattern. In their review, Lau et al. (submitted) divide the fMRI priming literature (N=13) to short and long SOA (stimulus onset asynchrony). In the priming literature, short SOA (250 ms) priming is assumed to depend on spreading activation while long SOA (1000 ms) is assumed to depend on strategic planning. In the context of the discussion here, short SOA reflects bottom-up pre-activation while long SOA reflects explicit prediction. Interestingly, a meta-analysis of the two priming types reveals a striking dissociation (cf. figure 3 in Lau et al. submitted).

Short SOA priming (or bottom-up pre-activation) was associated primarily with decreased activation in the middle temporal gyrus (MTG, BA 21). Long SOA priming
was associated with decreased activation in the IFC as well as in a number of loci within the temporal lobe (including the MTG). As pointed out by the authors, this spatial dissociation stands in contrast to the similar effect both priming types have on the N400 component. The authors propose that the specific brain distribution argues that the MTG is engaged in context independent lexical access while the IFC is engaged in context dependent integration.

The authors point to a further dissociation within the IFC. BA47 (orbital gyrus) in the ventral IFC exhibits decrease in activation for the related prime condition compared to the neutral baseline (paralleling the attenuated N400 response). BA 44/45 (Broca’s region) show an increase in activation for the unrelated prime condition compared to the neutral baseline. This suggests that activity in BA47 is sensitive to the predictability of the target in context but not to potential competition (same as with the N400), while activation in Broca’s region is sensitive to the competition between the item predicted by the unrelated prime and the actual target.

**The frequency effect**

The only fMRI study known to us to investigate the effects of word frequency on auditory word processing (Prabhakaran et al., 2006) has found increased activation in the IFC for low frequency words. Graves et al. (2007) also found activation in the IFC to be modulated by word frequency, using a picture naming task. Two papers looking at the effect of frequency on visual single word processing (Chee et al., 2003; Fiebach et al., 2002a) found low frequency-high frequency effect in the IFC as well. Though we have not conducted a proper meta-analysis, it appears that, across studies, BA 44 and 45 (Broca’s area) are the locus of the frequency effect in the IFC. Kronbichler et al. (2004) used a parametric design to investigate the effect of frequency. These authors report an effect of frequency in the left posterior MTG (as well as in the left IFC).

Summing up, fMRI investigation of pre-activation suggests two main loci of pre-activation effects, namely the left IFC and left MTG. It appears that pre-activation in the MTG is of the bottom-up type rather than a result of explicit prediction. Activat-
tion in the ventral IFC (BA 47) appears to be modulated by contextual predictability. Activation in Broca’s region in the IFC appears to be modulated by competition (indexed by increased activation for the distractor condition in long SOA priming and frequency effects).

2.6 The current experiment

Three recurring features of the empirical study of predictability have been a limiting factor in the field:

- Factorial design: most studies use a binary modulation of predictability (surprising versus non surprising). One rare exception is DeLong et al. (2005) who made use of a parametric design with 10 levels of cloze probability. The advantage of parametric design is that it is more powerful and can be used to test quantitative predictions and not only qualitative ones. (cf. also Demberg and Keller 2008; Vasishth et al. 2008)

- Sparse data: Predictability is estimated and tested for only one word per each sentence read or heard by the subject (usually the final one). This limitation arises from a number of sources. First, the use of a discrete factorial design requires comparison across similar contexts (as discussed above). Second, traditional calculation of cloze probability is highly time consuming. Third, the dependent measures used sometimes limit the number of observations per sentence.

- Decontextualization: Most experiments make use of single sentences or at most short paragraphs (2-3 sentences long).

Probabilistic Natural Language Processing tools can be made to produce quantitative surprisal or predictability estimates for every word in the sentence and can do so efficiently for large number of sentences. Brain imaging (as well as eye tracking) eliminates the need for an additional behavioral task (such as button press). As a
consequence, the prospects of a parametric, dense design using naturalistic language comprehension task become a reality.

Despite the clear interest in such designs, the use of computational tools to estimate predictability (cloze probability) or surprisal has been very limited. Stevens and Rumelhart (1975) trained an Augmented Transition Network (ATN) on a corpus and then used it to produce the cloze probabilities (or surprisal) of words in a short narrative. The authors found a high correlation between the predictions of the ATN and cloze probabilities computed in the traditional way as well as with word by word reading time.

Very recently, Demberg and Keller (2008) made use of Roark’s (2001) parser (the same parser used in the experiment reported here) to generate dense surprisal predictions for the Dundee corpus (Kennedy and Pynte, 2005), a large eye-tracking corpus of 50,000 words read by 10 subjects. The authors used a mixed linear model to control for low level dimensions of eye movement patterns and lexical properties such as length. The authors demonstrated a significant (though relatively small) effect of syntactic surprisal on fixation pattern during reading. A similar unpublished project in German has been reported in Vasishth et al. (2008). These authors also used trained probabilistic parsers to produce syntactic surprisal estimates which they then used to model eye-tracking data from Kliegl et al. (2004). Similarly to Demberg and Keller (2008), they found small but significant contribution for syntactic surprisal for both early and late eye movement measures. To our knowledge no similar attempt has been made using fMRI or electrophysiology.

In the current experiment we were interested to identify the neural correlates of pre-activation and expectation in the context of a naturalistic task. We were hoping to use the neural correlates of different aspects of expectation to evaluate the distinctions addressed above, namely:

- Distinguishing the effects of top-down expectation and bottom-up familiarity.
- Distinguishing lexical and syntactic surprisal
- Distinguishing syntactic surprisal and parsing complexity
Estimating the effect of entropy

bottom-up familiarity was quantified using log word frequency measures from the HAL corpus (Lund and Burgess, 1996), available through the English Lexicon Project web site (Balota et al., 2007). Top-down surprisal and number of derivation steps were calculated using the Roark parser (for a detailed description of this measure cf. ??). In order to avoid the confound of lexicality (open class words are less frequent and more surprising that closed class words) we restricted our analysis of frequency and lexical surprisal to open class items only. Finally, sentence level measures were obtained by averaging the word level values.

2.6.1 Correlation across the different measures

One important issue that needs to be addressed before moving on to the results is the question of correlation across our different measures. As we discussed in the introduction (1.5.6), all the measures were put into a single GL model and so any collinearity (or correlation) between them could potentially result in indeterminacy regarding the unique contribution of each of the measures. As we discussed in (2.3.3) syntactic surprisal and derivation steps are potentially closely related in a PCFG. Similarly, though frequency measures were not calculated in the same corpus as lexical surprisal (HAL and Brown respectively) it is possible that the former contributes significantly to the latter\textsuperscript{15}. In (2.1) we reproduce the correlation matrix of our 4 measures across the 4 narratives for open class words. We also included the length measure (the neural correlates of which have been discussed in the introduction):

As expected, syntactic surprisal and derivation length show a relatively high correlation ($> .34$). Syntactic surprisal shows no other important correlation. Lexical surprisal is correlated with both length and (negatively) with log frequency (log frequency goes up the more frequent, or ‘less surprising’, the word is). Finally length

\textsuperscript{15}Remember that lexical surprisal is based on the probability of the lexical insertion rule (e.g. \( V \rightarrow \text{eat} \)). At the limit, if the context contributes no information regarding the identity of the specific lexical item its probability will reduce to its probability given the Part of Speech Tag which is close, though not identical to frequency
and frequency are negatively correlated, as is usually the case in natural corpus (Zipf, 1949). Though these correlations appear important they are not too high to prevent a successful modeling in a GLM (Eric Zarahn, p.c.). However, variance that is not uniquely explained by a single measure (but may be ambiguous due to collinearity between the measures) will not be captured by the individual contrast.

In order to evaluate the shared contribution of the partially correlated measures we included in the results combined contrasts. The combined contrasts express the combined effects of the correlated measures including the shared effects. Naturally, these combined contrasts do not allow us to determine which of the participating measures is responsible for the particular distribution of activation.

An alternative would have been to orthogonalize the regressors, but that would have imposed an a-priori 'importance' scale which we had no grounds to apply (is length more basic than frequency or vice versa?). A third alternative would have been to extract components (as in Independent Component Analysis or Principle Component Analysis)\(^\text{16}\). The well known problem with such an approach regards the interpretation of the resulting components.

### 2.7 Results: Word class model

#### 2.7.1 Neural correlates of lexical surprisal and frequency

Peaks of positive correlation with lexical surprisal and peaks of negative correlation with frequency \((P < 0.001\) uncorrected, cluster threshold \(=20\)) are reported in tables

\(^{16}\text{cf. Hauk et al. (2006) for an approach along these lines for a lexical decision experiment.}\)
(2.2) and (2.3) respectively. Table (2.4) reports the peaks of positive and negative effect of the combined measure (lexical surprisal and frequency):

Table 2.2: Lexical surprisal results

<table>
<thead>
<tr>
<th>Region</th>
<th>Location</th>
<th>MNI coordinates (x,y,z)</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>L temporal</td>
<td>inferior temporal gyrus</td>
<td>-32 -46 -12</td>
<td>6.2577</td>
</tr>
<tr>
<td></td>
<td>posterior superior temporal sulcus</td>
<td>-40 -62 30</td>
<td>5.5165</td>
</tr>
<tr>
<td></td>
<td>inferior temporal gyrus</td>
<td>-50 -62 30</td>
<td>3.9982</td>
</tr>
<tr>
<td></td>
<td>middle temporal gyrus</td>
<td>-48 -56 -20</td>
<td>5.2748</td>
</tr>
<tr>
<td>L frontal</td>
<td>middle frontal gyrus</td>
<td>-42 24 46</td>
<td>6.124</td>
</tr>
<tr>
<td></td>
<td>middle frontal gyrus</td>
<td>-30 26 42</td>
<td>5.8612</td>
</tr>
<tr>
<td></td>
<td>middle frontal gyrus</td>
<td>-36 28 50</td>
<td>4.4618</td>
</tr>
<tr>
<td></td>
<td>superior frontal gyrus</td>
<td>-4 34 56</td>
<td>4.8289</td>
</tr>
<tr>
<td></td>
<td>superior frontal gyrus</td>
<td>-8 42 50</td>
<td>4.6335</td>
</tr>
<tr>
<td></td>
<td>superior frontal gyrus</td>
<td>-2 34 46</td>
<td>4.1057</td>
</tr>
<tr>
<td>L parietal</td>
<td>angular gyrus</td>
<td>-50 -56 18</td>
<td>5.9283</td>
</tr>
<tr>
<td></td>
<td>superior parietal gyrus</td>
<td>-58 -56 8</td>
<td>5.4089</td>
</tr>
<tr>
<td></td>
<td>precuneus</td>
<td>-36 -68 46</td>
<td>4.7447</td>
</tr>
<tr>
<td></td>
<td>lingual gyrus</td>
<td>-6 -86 -12</td>
<td>4.8529</td>
</tr>
<tr>
<td>L occipital</td>
<td>lingual gyrus</td>
<td>-16 -80 -16</td>
<td>4.6328</td>
</tr>
<tr>
<td>L occipital</td>
<td>middle occipital gyrus</td>
<td>-32 -82 26</td>
<td>4.6298</td>
</tr>
<tr>
<td>R temporal</td>
<td>occipito-temporal gyrus</td>
<td>36 -38 -12</td>
<td>5.238</td>
</tr>
<tr>
<td>R temporal</td>
<td>posterior superior temporal sulcus</td>
<td>40 -60 30</td>
<td>4.6141</td>
</tr>
<tr>
<td>R fusiform</td>
<td></td>
<td>28 -75 -22</td>
<td>5.6767</td>
</tr>
<tr>
<td>R frontal</td>
<td>superior frontal gyrus</td>
<td>10 44 48</td>
<td>4.4229</td>
</tr>
<tr>
<td>R occipital</td>
<td>inferior occipital gyrus</td>
<td>22 -88 -18</td>
<td>3.7237</td>
</tr>
<tr>
<td>R occipital</td>
<td>middle occipital sulcus</td>
<td>42 -84 10</td>
<td>4.898</td>
</tr>
</tbody>
</table>

Table 2.3: Log frequency results

<table>
<thead>
<tr>
<th>Region</th>
<th>Location</th>
<th>MNI coordinates (x,y,z)</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>L temporal</td>
<td>posterior middle temporal gyrus</td>
<td>-50 -62 -6</td>
<td>4.6554</td>
</tr>
<tr>
<td>L temporal</td>
<td>inferior temporal sulcus</td>
<td>-46 -50 -14</td>
<td>6.5404</td>
</tr>
<tr>
<td>L temporal</td>
<td>inferior temporal gyrus</td>
<td>-44 -46 -22</td>
<td>5.7885</td>
</tr>
<tr>
<td>L temporal</td>
<td>superior temporal gyrus</td>
<td>-56 -12 -4</td>
<td>6.5223</td>
</tr>
<tr>
<td>L frontal</td>
<td>inferior precentral sulcus</td>
<td>-46 42 22</td>
<td>4.3449</td>
</tr>
<tr>
<td>R temporal</td>
<td>anterior superior temporal sulcus</td>
<td>50 -6 -16</td>
<td>4.7542</td>
</tr>
<tr>
<td>R frontal</td>
<td>anterior frontal gyrus</td>
<td>8 52 28</td>
<td>4.825</td>
</tr>
</tbody>
</table>

The combined measure (lexical surprisal and negative log frequency) positively correlated with activity in regions within the left temporal lobe (superior temporal sulcus and gyrus, inferior temporal sulcus), the temporal parietal junction (in particular the angular gyrus) and left inferior frontal cortex (opercular and triangular parts). Correlated activation in the right hemisphere was less robust but localized in similar regions. Frequency effects not overlapping the effect of the combined measure were limited to the left superior temporal gyrus and right anterior superior temporal sulcus. Finally, unique effects of lexical surprisal were observed in the left middle
Table 2.4: Combined lexical surprisal and frequency results

### Positive effect of combined lexical surprisal and frequency

<table>
<thead>
<tr>
<th>Region</th>
<th>Location</th>
<th>MNI coordinates (x,y,z)</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>L temporal</td>
<td>superior temporal sulcus</td>
<td>-48 -56 8</td>
<td>7.687</td>
</tr>
<tr>
<td></td>
<td>superior temporal sulcus</td>
<td>-50 -48 12</td>
<td>7.4551</td>
</tr>
<tr>
<td></td>
<td>occipital temporal sulcus</td>
<td>-48 -56 -20</td>
<td>7.5278</td>
</tr>
<tr>
<td></td>
<td>inferior temporal sulcus</td>
<td>-48 -48 -14</td>
<td>5.7262</td>
</tr>
<tr>
<td></td>
<td>inferior temporal sulcus</td>
<td>-42 -54 -10</td>
<td>4.3855</td>
</tr>
<tr>
<td></td>
<td>superior temporal gyrus</td>
<td>-56 -12 -6</td>
<td>4.6242</td>
</tr>
<tr>
<td></td>
<td>superior temporal gyrus</td>
<td>-54 -8 -14</td>
<td>3.966</td>
</tr>
<tr>
<td></td>
<td>inferior temporal sulcus</td>
<td>-40 -36 -20</td>
<td>4.5841</td>
</tr>
<tr>
<td>L frontal</td>
<td>superior temporal sulcus</td>
<td>-50 -48 12</td>
<td>7.4551</td>
</tr>
<tr>
<td></td>
<td>occipital temporal sulcus</td>
<td>-48 -56 -20</td>
<td>7.5278</td>
</tr>
<tr>
<td></td>
<td>inferor temporal sulcus</td>
<td>-48 -48 -14</td>
<td>5.7262</td>
</tr>
<tr>
<td></td>
<td>inferor temporal sulcus</td>
<td>-42 -54 -10</td>
<td>4.3855</td>
</tr>
<tr>
<td></td>
<td>superior temporal gyrus</td>
<td>-56 -12 -6</td>
<td>4.6242</td>
</tr>
<tr>
<td></td>
<td>superior temporal gyrus</td>
<td>-54 -8 -14</td>
<td>3.966</td>
</tr>
<tr>
<td></td>
<td>inferior temporal sulcus</td>
<td>-40 -36 -20</td>
<td>4.5841</td>
</tr>
<tr>
<td>L parietal</td>
<td>angular gyrus</td>
<td>-56 -52 -26</td>
<td>5.1325</td>
</tr>
<tr>
<td></td>
<td>angular gyrus</td>
<td>-44 -70 -28</td>
<td>4.3307</td>
</tr>
<tr>
<td></td>
<td>inferor temporal sulcus</td>
<td>-48 -10 -28</td>
<td>5.6589</td>
</tr>
<tr>
<td></td>
<td>inferor temporal sulcus</td>
<td>-38 -16 -28</td>
<td>4.8913</td>
</tr>
<tr>
<td></td>
<td>inferor precentral sulcus</td>
<td>-38 4 32</td>
<td>4.6264</td>
</tr>
<tr>
<td>R parietal</td>
<td>angular gyrus</td>
<td>38 -36 -24</td>
<td>4.9837</td>
</tr>
<tr>
<td>R temporal</td>
<td>posterior superior temporal sulcus</td>
<td>40 -60 -20</td>
<td>4.7712</td>
</tr>
<tr>
<td></td>
<td>superior temporal sulcus</td>
<td>54 -52 -12</td>
<td>4.5161</td>
</tr>
<tr>
<td>R fusiform</td>
<td>anterior superior frontal gyrus</td>
<td>12 52 -20</td>
<td>6.9869</td>
</tr>
<tr>
<td></td>
<td>paracentral sulcus</td>
<td>8 -20 62</td>
<td>5.0049</td>
</tr>
<tr>
<td>R frontal</td>
<td>inferior frontal sulcus</td>
<td>42 16 -26</td>
<td>4.8746</td>
</tr>
<tr>
<td></td>
<td>angular gyrus</td>
<td>8 -15 34</td>
<td>4.6902</td>
</tr>
<tr>
<td>R parietal</td>
<td>cingulate gyrus</td>
<td>42 -54 52</td>
<td>4.9585</td>
</tr>
<tr>
<td>R cingulate</td>
<td>posterior cingulate gyrus</td>
<td>42 -54 52</td>
<td>4.9585</td>
</tr>
<tr>
<td>R cingulate</td>
<td>intraparietal sulcus</td>
<td>-4 -32 32</td>
<td>3.9013</td>
</tr>
<tr>
<td>R cingulate</td>
<td>cingulate gyrus</td>
<td>6 -38 30</td>
<td>5.937</td>
</tr>
</tbody>
</table>

### Negative effect of combined lexical surprisal and frequency

<table>
<thead>
<tr>
<th>Region</th>
<th>Location</th>
<th>MNI coordinates (x,y,z)</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>L cingulate</td>
<td>posterior cingulate gyrus</td>
<td>-4 -32 32</td>
<td>3.9013</td>
</tr>
<tr>
<td>R parietal</td>
<td>intraparietal sulcus</td>
<td>42 -54 52</td>
<td>4.9585</td>
</tr>
<tr>
<td>R cingulate</td>
<td>cingulate gyrus</td>
<td>6 -38 30</td>
<td>5.937</td>
</tr>
</tbody>
</table>

The left superior frontal gyrus, the left angular and superior parietal gyre and in the occipital lobe (in particular the angular gyrus).
Figure 2-1: **A:** Positive correlation with lexical surprisal (orange clusters in the angular gyrus and dorsal prefrontal cortex), negative correlation with frequency (red clusters along the superior temporal lobe, posterior temporal lobe and inferior frontal gyrus) and the combined effect (yellow clusters in inferior frontal gyrus, posterior and inferior temporal lobe) in the left hemisphere. **B:** A ventral view of positive correlation with lexical surprisal (orange), negative correlation with frequency (red) and the combined effect (yellow) in and around the Word Form Area in the left hemisphere fusiform gyrus. **C:** Positive correlation with lexical surprisal (orange cluster in the angular gyrus), negative correlation with frequency (red cluster in the anterior temporal lobe) and the combined effect (yellow clusters in posterior temporal and frontal lobes) in the right hemisphere. All data presented on a two dimensional representation of an 'inflated' cortex. The group t map was thresholded at $p < 0.001$, uncorrected, cluster threshold=20.
2.7.2 Neural correlates of syntactic surprisal and number of derivation steps

Peaks of positive correlation with number of derivation steps and syntactic surprisal measures ($P < 0.001$ uncorrected, cluster threshold =20) are reported in tables (2.5) and (2.6) respectively. Table (2.7) reports the peaks of positive and negative effect of the combined measure (syntactic surprisal and number of derivation steps):

Table 2.5: Derivation steps results

<table>
<thead>
<tr>
<th>Region</th>
<th>Location</th>
<th>MNI coordinates (x,y,z)</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>L parietal</td>
<td>angular gyrus</td>
<td>-52 -56 24</td>
<td>4.9885</td>
</tr>
<tr>
<td></td>
<td>angular gyrus</td>
<td>-54 -64 30</td>
<td>4.473</td>
</tr>
<tr>
<td></td>
<td>angular gyrus</td>
<td>-62 -56 22</td>
<td>4.3342</td>
</tr>
<tr>
<td>R temporal</td>
<td>anterior mid. temporal gyrus</td>
<td>52 0 -40</td>
<td>8.2167</td>
</tr>
<tr>
<td></td>
<td>posterior superior temporal sulcus</td>
<td>54 -48 28</td>
<td>4.1649</td>
</tr>
<tr>
<td></td>
<td>superior temporal sulcus</td>
<td>46 -26 0</td>
<td>6.3239</td>
</tr>
<tr>
<td></td>
<td>posterior superior temporal sulcus</td>
<td>60 -4 14</td>
<td>4.7215</td>
</tr>
<tr>
<td></td>
<td>anterior superior temporal sulcus</td>
<td>54 -12 12</td>
<td>4.6937</td>
</tr>
<tr>
<td></td>
<td>temporal pole</td>
<td>46 16 -40</td>
<td>5.031</td>
</tr>
<tr>
<td>R frontal</td>
<td>superior temporal gyrus</td>
<td>66 -34 4</td>
<td>4.3413</td>
</tr>
<tr>
<td>R parietal</td>
<td>anterior superior frontal gyrus</td>
<td>8 54 20</td>
<td>5.0285</td>
</tr>
<tr>
<td></td>
<td>sub-parietal sulcus</td>
<td>2 -56 32</td>
<td>6.3793</td>
</tr>
<tr>
<td></td>
<td>sub-parietal sulcus</td>
<td>8 -50 22</td>
<td>5.9665</td>
</tr>
<tr>
<td></td>
<td>precuneus</td>
<td>4 -60 40</td>
<td>5.6311</td>
</tr>
<tr>
<td></td>
<td>angular gyrus</td>
<td>48 -56 26</td>
<td>5.1422</td>
</tr>
</tbody>
</table>

Table 2.6: Syntactic surprisal results

<table>
<thead>
<tr>
<th>Region</th>
<th>Location</th>
<th>MNI coordinates (x,y,z)</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>R frontal</td>
<td>middle frontal gyrus</td>
<td>50 40 26</td>
<td>5.6074</td>
</tr>
<tr>
<td>R parietal</td>
<td>angular gyrus</td>
<td>40 -78 44</td>
<td>5.2519</td>
</tr>
<tr>
<td></td>
<td>angular gyrus</td>
<td>42 -60 50</td>
<td>4.5188</td>
</tr>
</tbody>
</table>

In contrast to lexical surprisal, the effect of syntactic surprisal and derivation steps was right lateralized. In the left hemisphere there was no unique contribution of syntactic surprisal. The unique contribution of derivation steps in the left hemisphere overlapped with the combined effect in the angular gyrus and the precuneus. In the right hemisphere, syntactic surprisal had a unique effect in the parietal lobe and the medial frontal cortex. Unique contribution of derivation steps was observed along the superior temporal sulcus and the temporal pole. The combined measure had a significant effect in the angular gyrus and the temporal parietal junction as well as anterior prefrontal cortex and the precuneus.
Table 2.7: Derivation steps and syntactic surprisal combined

Positive correlation with syntactic surprisal and number of derivation steps: open class words

<table>
<thead>
<tr>
<th>Region</th>
<th>Location</th>
<th>MNI coordinates (x,y,z)</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>L temporal</td>
<td>anterior mid temporal gyrus</td>
<td>-54 -4 -28</td>
<td>7.8385</td>
</tr>
<tr>
<td></td>
<td>anterior mid temporal gyrus</td>
<td>-50 -6 -32</td>
<td>4.6467</td>
</tr>
<tr>
<td></td>
<td>temporal pole</td>
<td>-46 10 -40</td>
<td>4.5304</td>
</tr>
<tr>
<td></td>
<td>posterior superior temporal gyrus</td>
<td>-44 -46 28</td>
<td>5.5663</td>
</tr>
<tr>
<td></td>
<td>superior temporal sulcus</td>
<td>-46 -40 0</td>
<td>5.7331</td>
</tr>
<tr>
<td></td>
<td>middle temporal gyrus</td>
<td>-68 -36 0</td>
<td>5.5704</td>
</tr>
<tr>
<td>L parietal</td>
<td>angular gyrus</td>
<td>-50 -56 24</td>
<td>4.881</td>
</tr>
<tr>
<td></td>
<td>angular gyrus</td>
<td>-42 -50 26</td>
<td>4.5802</td>
</tr>
<tr>
<td></td>
<td>sub-parietal sulcus/precuneus</td>
<td>-10 -50 42</td>
<td>4.1254</td>
</tr>
<tr>
<td>R temporal</td>
<td>anterior inferior temporal gyrus</td>
<td>52 0 -40</td>
<td>8.096</td>
</tr>
<tr>
<td></td>
<td>superior temporal sulcus</td>
<td>60 -4 -14</td>
<td>5.7152</td>
</tr>
<tr>
<td></td>
<td>middle temporal gyrus</td>
<td>50 -16 -14</td>
<td>4.7968</td>
</tr>
<tr>
<td>R temporal</td>
<td>temporal pole</td>
<td>42 14 -42</td>
<td>5.1813</td>
</tr>
<tr>
<td></td>
<td>temporal pole</td>
<td>40 12 -30</td>
<td>4.2608</td>
</tr>
<tr>
<td>R frontal</td>
<td>anterior superior frontal gyrus</td>
<td>16 66 20</td>
<td>5.6921</td>
</tr>
<tr>
<td></td>
<td>anterior superior frontal gyrus</td>
<td>6 52 26</td>
<td>5.679</td>
</tr>
<tr>
<td></td>
<td>anterior superior frontal gyrus</td>
<td>0 52 36</td>
<td>4.5198</td>
</tr>
<tr>
<td></td>
<td>anterior superior frontal gyrus</td>
<td>6 46 44</td>
<td>4.3965</td>
</tr>
<tr>
<td>R parietal</td>
<td>angular gyrus</td>
<td>50 -60 20</td>
<td>7.59</td>
</tr>
<tr>
<td></td>
<td>sub-parietal sulcus/precuneus</td>
<td>4 -58 36</td>
<td>5.1129</td>
</tr>
<tr>
<td></td>
<td>sub-parietal sulcus/precuneus</td>
<td>6 -50 34</td>
<td>4.6634</td>
</tr>
<tr>
<td>R occipital</td>
<td>middle occipital gyrus</td>
<td>48 -64 32</td>
<td>5.6896</td>
</tr>
</tbody>
</table>

Negative correlation with syntactic surprisal and number of derivation steps: open class words

<table>
<thead>
<tr>
<th>Region</th>
<th>Location</th>
<th>MNI coordinates (x,y,z)</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>L temporal</td>
<td>occipital temporal sulcus</td>
<td>-54 -60 -8</td>
<td>5.1226</td>
</tr>
<tr>
<td>L frontal</td>
<td>triangular gyrus</td>
<td>-42 28 10</td>
<td>477663</td>
</tr>
</tbody>
</table>
Figure 2-2:  

A: Positive correlation with number of derivation steps (orange cluster in the angular gyrus) and the combined effect (yellow clusters in the TPJ, mid temporal lobe and temporal pole) in the left hemisphere.  

B: A medial view of positive correlation with number of derivation steps (orange cluster in the precuneus) and the combined effect (yellow cluster in the precuneus) in the left hemisphere.  

C: Positive correlation with number of derivation steps (orange clusters in the TPJ and along the temporal lobe), syntactic surprisal (red cluster in the parietal lobe) and the combined effect (yellow clusters in the TPJ, temporal lob and superior frontal gyrus) in the right hemisphere.  

D: A medial view of positive correlation with number of derivation steps (orange clusters in the precuneus and anterior prefrontal cortex) and the combined effect (yellow clusters in the precuneus and anterior prefrontal cortex) in the right hemisphere.  

All data presented on a two dimensional representation of an 'inflated' cortex. The group t map was thresholded at $p < 0.001$, uncorrected, cluster threshold=20.
2.8 Discussion: Word class model

2.8.1 The neural correlates of lexical surprise

The left lateralization of the effect of lexical pre-activation is in line with the general tendency of language related processing to be left lateralized and with other imaging studies of the effect of contextual congruency (cf. Lau et al. submitted for a review) and frequency (Prabhakaran et al., 2006).

Broca's area in the left inferior frontal gyrus was activated by the combined contrast. This region has been consistently implicated by word frequency manipulations (Chee et al., 2003; Fiebach et al., 2002a; Prabhakaran et al., 2006; Graves et al., 2007). BA 45 (the triangularis part of Broca's area) has been implicated by multiple experiments manipulating contextual facilitation of lexical access (10 out of the 20 studies reported in Lau et al. submitted).

Wagner et al. (2001) argue that BA 45 is specifically engaged in controlled semantic retrieval. BA 44 (the opercular part of Broca's area) has been also observed to participate in semantic and lexical retrieval tasks. Vigneau et al. (2006), based on a large scale meta-analysis, propose that this region is part of the working memory network for semantic information. Putting together the perspective emerging from the memory literature with the linguistic literature and the results here we suggest (cf. also Lau et al. submitted for a similar perspective) that Broca's area is specifically engaged in mediating competition during lexical retrieval and/or integration of the lexical item into the current sentential frame.

The posterior middle temporal gyrus, which was activated by both frequency and the combined measure has been long known to subserve auditory comprehension. Though it is not strictly a part of Wernicke's area (which usually is restricted to the posterior superior temporal gyrus; Wise et al. 2001) it has been implicated in lexical processing by both functional imaging experiments (Rissman et al., 2003; Rodd et al., 2005) and lesion studies (Dronkers et al., 2004). Lau et al. (submitted), in a recent

\[17\text{Covering both frequency and lexical surprisal}\]
meta-analysis of fMRI studies of semantic priming and semantic anomaly in sentential context, found the posterior middle temporal region to be particularly affected by priming (8 of 13 studies) but not by sentential context anomaly (only 1 out of 16 studies).

Dronkers and colleagues performed a large scale voxel-lesion symptom mapping (Bates et al., 2003) with 64 left hemisphere stroke patients and a battery of lexical and syntactic tests. Patients with a lesion in the posterior middle temporal region were found to be the worst performing sub-group, with severe deficits already in word level tasks (and so naturally also impaired on sentence level tasks that presuppose lexical processing). In fact, lesions in Wernicke’s area proper did not produce such dramatic deficits, leading the authors to suggest that it is damage to the posterior middle temporal region that is responsible to Wernicke type aphasia.

In Lau et al.’s (submitted) meta-analysis of priming studies, the MTG is observed to be the primary region to be affected by short SOA priming. Short SOA priming is generally assumed to be mediated via spreading activation (or bottom-up pre-activation) rather than strategic prediction.

While most fMRI studies of frequency effect do not report activation in this region, this could be due to the use of a binary contrast (low versus high frequency) that might not be sensitive enough. Interestingly, Kronbichler et al. (2004), using a parametric design to investigate the effect of frequency on single word reading, report negative correlation with frequency in the MTG (as well as, as in other studies, the left IFC).

Taken together, these findings suggest a central role for this region in lexical access and in particular aspects of lexical access (presumably early) that are not influenced by prediction but are sensitive to the level of pre-activation. The fact that in our study activity in this region was modulated more by frequency than semantic surprisal lends further support to this conclusion. We then propose that brain activation in the MTG supports the bottom-up view of lexical access.

The mid superior temporal gyrus and sulcus present a similar picture. In these regions it is also the effect of frequency rather than lexical surprisal that appears significant. Activation in these regions is often associated with phonological processing
Hickok and Poeppel (2007) elaborate a model of speech perception based on current neurolinguistic findings. In their model, the mid superior temporal sulcus and adjacent regions support sublexical phonological processing and the formation, or retrieval, of the phonological word form en-route to more posterior regions within the middle temporal gyrus where the resulting phonological representations are mapped onto semantic word meanings.

Okada and Hickok (2006) demonstrated that activity in this region during auditory lexical access is modulated by phonological neighborhood size. More specifically, these authors found items with larger phonological neighborhoods to produce more activation in this region compared to words with small neighborhoods. This fact that activation in this region is sensitive to competition among phonologically similar forms (as in the cohort model, Marslen-Wilson 1987) suggests a ‘bottom-up’ view of the role of this region since this type of competition is not dependent on prior context, but instead, properties of the incoming stimulus.

The fact that contextual expectedness (surprisal) does not significantly affect activation in this region lends further support to the view that sublexical phonological processing is facilitated by familiarity but not contextual information or predictability. This speculative conclusion is in line with bottom-up approaches to phonological and early lexical processing (Cutler and Norris, 1979).

Activation in the angular gyrus presents the mirror image, with lexical surprisal, rather than frequency, exhibiting a significant effect. The angular gyrus has been traditionally associated, following Geschwind, with reading and orthography to phonology mapping (Horwitz et al., 1998). However, recent imaging studies have found this region to be involved in auditory language processing as well (Humphries et al., 2006). In effect, in the model elaborated by Vigneau and colleagues, the angular gyrus is part of the same a-modal semantic network that BA 44 is. The fact that only lexical surprisal has a unique contribution in this region suggests that the semantic processing taking place in the angular gyrus is sensitive to the specific contextual (or sentential)

The size of the phonological neighborhood of a word is determined by the number of words that can be formed by replacing any one phoneme in the word.
constraints rather than to a general familiarity with the specific word.

Humphries et al. (2006) found the angular gyrus to be more active to auditory presented sentences than to scrambled word lists and Ramachandran found that the left angular gyrus is critical for the interpretation of metaphoric meaning (cited in Giora 2007). Put together with our finding that activity in this region exhibits sensitivity to contextual constraints, this evidence suggests a refinement of Vigneau and colleagues' model to the effect that the angular gyrus specifically supports sentence level (or compositional) aspects of semantic processing. This conclusion receives further support from the lesion study by Dronkers et al. (2004) who found lesions in the angular gyrus to be correlated with performance decrease on sentential level tasks rather than word level ones. We propose that the angular gyrus subserves integration of the retrieved lexical item in the larger sentential structure.

Two subregions within left Brodmann's area 8 (in the prefrontal cortex) also showed surprisal but no frequency effect. Though BA 8 is usually not discussed in the context of language processing, its medial portion has been recently implicated by a number of experiments exploring the effect of uncertainty (Volz et al., 2005; Ridderinkhof et al., 2004). Volz et al parametrically varied the level of uncertainty during decision making in two different experiments (in each experiment the source of uncertainty was different). They found the medial portion of BA 8 to be parametrically activated by uncertainty across both experiments.

Ridderinkhof et al. (2004) produced a meta-analysis of the correlates of cognitive control in the medial frontal cortex. Volz and colleagues note that of the 38 studies grouped in that meta-analysis, only 5 varied decision uncertainty while most other studies varied response uncertainty\(^{19}\). All 5 studies which manipulated decision uncertainty reported relevant activation in BA 8 (response uncertainty modulated activation in the anterior and dorsal cingulate). In this respect, the variation in the amount of available contextual information to inform prediction regarding upcoming

\(^{19}\)One specific characteristic of decision uncertainty is that the uncertainty is not the consequence of competition between a correct response and a distractor but is due to lack of full knowledge regarding which is the correct response. See Volz et al for a detailed definition of these terms.
lexical items (and as a consequence, the extent of lexical surprisal) can be conceptualized as a case of decision uncertainty with the incoming lexical item as the feedback. Though we have not quantified directly decision uncertainty, there is a general relation between surprise and entropy (Strange et al., 2005) where entropy can be thought of as the running average of surprise (cf. 2.4).

Finally, both frequency and lexical surprisal had a significant effect on activation in the left posterior inferior temporal lobe (2-1.B). These clusters of activation, (MNI coordinates [-46, -50, -14] and [-48, -56, -20] respectively) are very close to the center of a region known as the Word Form Area (Cohen et al., 2002) ([ -44, -58, -15] as defined by the meta-analysis in Jobard et al. 2003). This region has been consistently implicated by word and (certain) pseudo-word reading tasks and has been considered as the putative site for visual word processing by Cohen and colleagues as well as many other researchers (cf. Jobard et al. 2003 for a review). A number of studies of the effect of word frequency on visual word processing report activation in this region (Kronbichler et al., 2004; Hauk et al., 2008).

We are aware of only one other paper in the literature (Price et al., 2003) that has reported activation in this region for an auditory task (auditory word repetition). However the overt production and metalinguistic task make it difficult to interpret these results (cf. Dehaene et al. 2002). Despite Price and colleagues' evidence against the strictly unimodal character of this region, theoretical and empirical work on this region have focused exclusively on visual (sub)lexical processing (Dehaene et al., 2005).

The fact reported here, that activity in this same region is modulated by frequency as well as contextual surprisal of auditorily presented words suggests a different view of the role of this region in language processing. In particular, it suggests a view of this region as supporting a-modal lexical storage. In this respect it is worthwhile noting that according to Hickok and Poeppel (2007) (and others), the inferior temporal

---

It is possible, of course, to attempt to maintain the unimodal view of this region by claiming that the processing of less frequent or less expected words requires (or results in) access to their orthographic form and, as a consequence, activation of the word form area (David Caplan, p.c.).
region supports lexical semantic retrieval and more generally stores object specific knowledge (Martin and Chao, 2001; Tyler et al., 2004).

2.8.2 The neural correlates of syntactic surprisal and derivation steps

The unique contribution of number of derivation steps was overall stronger than that of syntactic surprisal. In fact, no unique contribution of syntactic surprisal exceeded the threshold in the left hemisphere. The shared effect had a bilateral distribution with clusters of activation in the anterior temporal lobe, the mid temporal gyrus and the angular gyrus (or temporal parietal junction) as well as the posterior precuneus. Right lateralized effects were observed in the mid frontal gyrus.

The anterior temporal lobe has been traditionally associated with word level semantic information in particular due to the relation between atrophy in this region and dementia or loss of semantic knowledge. However, over the past few years this region has become increasingly implicated in syntactic (or sentential) processing (Jobard et al., 2007; Foki et al., 2008; Humphries et al., 2006; Vandenberghe et al., 2002; Hickok and Poeppel, 2007; Vigneau et al., 2006; Lau et al., submitted). In line with this emerging view, we found this region to be sensitive to syntactic rather than lexical factors.

The activation in the temporal pole is ambiguous between number of derivation steps and syntactic surprisal. Certain authors (Friederici and von Cramon, 2001) are of the view that anterior temporal regions subserve some sort of basic syntactic computation which is not sensitive to the complexity of the structure (maybe morphosyntax, as suggested by Dronkers et al. 1994). Under this view number of parser steps might not be expected to produce increased activation in this region. At the same time, as discussed above (2.5.3), Friederici proposes that the ELAN’s source is in the temporal pole. Under the perspective advanced by Lau et al. (2006) that the ELAN reflects syntactic prediction violation, we could suggest that the activation in this region seen here is, in effect, due to syntactic surprisal. Data from the Sentence
model to be discussed later supports this view.

Indefrey and Levelt (2004) conducted a meta-analysis of over 80 neuroimaging and electrophysiological studies of word production and perception. Based on this analysis they suggested that lemma information (or the syntactic features of the lexical items, in contrast to conceptual or semantic ones) is stored or processed in the mid portion of the middle temporal gyrus. Our finding of syntactic surprisal related activation in this region is in agreement with this view. However in our study this activation was bilateral while Indefrey and Levelt report a left lateralized pattern. The difference may be due to the fact that the meta-analysis grouped experiments looking at single word processing while our experiment involved lexical access in sentential context.

The emerging view of the role of the angular gyrus in sentence level processing has already been pointed out in the previous section. It is interesting to note that while the syntactic and lexical effects partially overlap in the left angular gyrus, the right angular gyrus exhibits only syntactic effects. Furthermore, in the left hemisphere, the syntactic effect has a more posterior-dorsal distribution extending further into the parietal lobe and the supra-marginal gyrus while the lexical effect extends ventrally into mid temporal regions.

To our knowledge, there is no published discussion of functional segregation within the angular gyrus (at least in the context of language processing). However, this difference in distribution is in line with the conclusions drawn by Ben Shalom and Poeppel (2008) from a survey of current neurolinguistic models. The authors draw a very coarse distinction between the role of parietal regions in analysis versus temporal regions in memorizing. Taking into consideration our results regarding the effect of structural ambiguity and long distance dependencies in the parietal cortex (to be reviewed in the following chapters) , it is suggestive to think of the angular gyrus as an interface between a temporal lexical network and a parietal-frontal syntactic network.

The function of the precuneus, a posterio-medial subdivision of the parietal lobe has remained largely unexplored until recently. The main reason for this state of affairs was that lesions to this region are rare. With the advent of neuroimaging, the
dependence of neuro-cognitive science on lesion studies has been attenuated and as a consequence the precuneus has been emerging as a key player in a diverse set of cognitive states and functions such as episodic and autobiographical memory, imagery, theory of mind and self reflection cf. Cavanna and Trimble 2006; Spreng et al. 2008 for recent reviews). Data from primate studies, functional imaging of cognitive tasks (Cavanna and Trimble, 2006) as well as a recent resting state functional connectivity study (Marguiles et al., 2008) strongly indicate that the precuneus is functionally heterogeneous.

One emerging gross division within the precuneus distinguishes between the anterior and posterior precuneus. The anterior part has been repeatedly implicated in self referential tasks as well as part of the ‘default network’ (Raichle et al., 2001; Buckner et al., 2008). The posterior part has been found active, among others, in episodic memory retrieval, mental imagery, theory of mind tasks and navigation. While earlier work associated precuneus activation in episodic memory to mental imagery (Fletcher et al., 1995), more recent studies found it to be active in non-imagery related episodic memory of musical sequences and abstract words (Cavanna and Trimble, 2006). Cabeza et al. (2003) investigated the neural correlates of different stages of episodic memory retrieval. They found the posterior precuneus to be particularly engaged in the post recall stage. They proposed that the precuneus plays a role in the 'processing of internally generated stimuli' (p. 390). Elaborating on Cabeza’s and colleagues proposal one could view the posterior precuneus as supporting the representation or processing of hierarchically complex internal representations. Music, theory of mind, navigation and associative or episodic memory all require complex internal representations and so we can identify a unified role for the posterior precuneus throughout these diverse domains.

Given this view of the posterior precuneus it would be natural to expect it to have a role in syntactic processing, which crucially involves internally generated hierarchical representations. Contrary to this expectation, the posterior precuneus has not been considered as part of the syntactic processing network. However this could be the result of the long lasting dominance of lesion based models in neurolinguistics. Many
imaging studies looking at the effects of syntactic complexity focused on Broca’s area in the inferior frontal gyrus and Wernicke’s area in the posterior temporal lobe (for example Just et al. 1996; Ben-Shachar et al. 2003; Santi and Grodzinsky 2007).

(Shetreet et al., 2007) used a parametric design to investigate the neural correlates of number of complements of a verb (experiment 1). The authors compared unergative sentences (zero complements), transitive verbs (one complement) and ditransitive verbs (two complements). To avoid a length confound, the authors added post verbal adjuncts to make all sentences of the same length. Under the assumption (shared by the authors) that richer argument structure implies a more articulated VP structure (Larson, 1988), the resulting design instantiated a parametric manipulation of hierarchical complexity. The authors found significant effects only in the right posterior precuneus and anterior cingulate. Due to the adherence by the authors to the classic lesion model, they considered the precuneus activation to be linguistically irrelevant and, in fact, proceeded to conclude that number of arguments of a verb does not have neurolinguistic consequences (‘It thus seems that when a verb is accessed, the number of branches that should be constructed within the same tree does not tax linguistic resources’ p. 1967).

We suggest, alternatively, that the graded activation observed by the authors in the posterior precuneus reflects the gradual increase in syntactic complexity across conditions. In line with this view of the role of the posterior precuneus, we found a large unique contribution to activation in that region from the number of parser steps (but not syntactic surprisal). Quite intuitively (and easily verifiable by examination of the text), number of parser steps increases with the increase in syntactic complexity (such as relative clauses, sentential complements, small clauses and the like). The lack of a unique contribution from syntactic surprisal in this region suggests that this region is not engaged in prediction or conflict monitoring.
2.9 Results: The Sentence model

2.9.1 Neural correlates of lexical entropy and averaged frequency

Peaks of positive correlation with lexical entropy and peaks of negative correlation with averaged frequency ($P < 0.001$ uncorrected, cluster threshold =20) are reported in tables (2.8) and (2.9) respectively. Table (2.10) reports the peaks of the positive effect of the combined measure (lexical entropy and averaged frequency):

<table>
<thead>
<tr>
<th>Positive correlation with lexical entropy</th>
<th>MNI coordinates (x,y,z)</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>L temporal occipito-temporal sulcus</td>
<td>-34 -40 -12</td>
<td>5.7508</td>
</tr>
<tr>
<td>L temporal fusiform</td>
<td>-36 -42 -20</td>
<td>4.3881</td>
</tr>
<tr>
<td>L temporal middle frontal gyrus</td>
<td>-34 30 48</td>
<td>5.742</td>
</tr>
<tr>
<td>L temporal superior frontal gyrus</td>
<td>-12 42 52</td>
<td>4.7501</td>
</tr>
<tr>
<td>L frontal orbital gyrus</td>
<td>-30 38 -8</td>
<td>4.2478</td>
</tr>
<tr>
<td>L frontal orbital sulcus</td>
<td>-36 32 -16</td>
<td>3.6832</td>
</tr>
<tr>
<td>L frontal fusiform</td>
<td>34 -40 -16</td>
<td>4.4547</td>
</tr>
<tr>
<td>L frontal occipito-temporal sulcus</td>
<td>42 -36 -14</td>
<td>4.0243</td>
</tr>
<tr>
<td>R temporal occipito-temporal sulcus</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Negative correlation with averaged frequency</th>
<th>MNI coordinates (x,y,z)</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>L temporal occipital-temporal sulcus</td>
<td>-48 -48 -14</td>
<td>5.2964</td>
</tr>
<tr>
<td>L temporal inferior temporal sulcus</td>
<td>-46 -54 -6</td>
<td>4.4149</td>
</tr>
<tr>
<td>L frontal inferior frontal sulcus</td>
<td>-42 20 28</td>
<td>6.073</td>
</tr>
<tr>
<td>L frontal IFG: triangularis part</td>
<td>-50 24 18</td>
<td>5.8813</td>
</tr>
<tr>
<td>L frontal precentral sulcus</td>
<td>-42 4 30</td>
<td>5.7497</td>
</tr>
<tr>
<td>R frontal inferior frontal sulcus</td>
<td>42 26 20</td>
<td>6.3136</td>
</tr>
<tr>
<td>R frontal inferior frontal sulcus</td>
<td>46 18 20</td>
<td>5.0085</td>
</tr>
</tbody>
</table>

The effect of the two lexical factors and their combination was strongly left lateralized (as a consequence we rendered only the results for the left hemisphere in 2-3). Averaged lexical frequency had a unique contribution to activation in the inferior temporal gyrus and inferior temporal gyrus. Unique contribution of lexical entropy was observed in the dorsal prefrontal cortex, the ventral frontal cortex (orbital region) and the fusiform region in the temporal lobe. The combined measure had a significant effect in the inferior frontal gyrus, the posterior middle temporal lobe and the fusiform gyrus.
Table 2.10: Combined sentence entropy and averaged frequency measure

<table>
<thead>
<tr>
<th>Region</th>
<th>Location</th>
<th>MNI coordinates (x,y,z)</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>L temporal</strong></td>
<td>inferior temporal gyrus</td>
<td>-46 -52 -14</td>
<td>6.7863</td>
</tr>
<tr>
<td></td>
<td>posterior middle temporal gyrus</td>
<td>-48 -56 6</td>
<td>6.4965</td>
</tr>
<tr>
<td></td>
<td>inferior temporal sulcus</td>
<td>-52 -64 -2</td>
<td>5.806</td>
</tr>
<tr>
<td></td>
<td>fusiform gyrus</td>
<td>-30 -36 -18</td>
<td>4.4455</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-30 -46 -14</td>
<td>3.9655</td>
</tr>
<tr>
<td><strong>L frontal</strong></td>
<td>precentral sulcus</td>
<td>-48 6 30</td>
<td>9.5634</td>
</tr>
<tr>
<td></td>
<td>inferior frontal sulcus</td>
<td>-44 18 30</td>
<td>4.1441</td>
</tr>
<tr>
<td></td>
<td>IFG: triangularis part</td>
<td>-48 34 6</td>
<td>6.3754</td>
</tr>
<tr>
<td></td>
<td>inferior frontal sulcus</td>
<td>-50 26 18</td>
<td>4.5217</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-42 26 28</td>
<td>3.6871</td>
</tr>
<tr>
<td></td>
<td>orbital sulcus</td>
<td>-22 30 -18</td>
<td>5.5137</td>
</tr>
<tr>
<td></td>
<td>superior frontal gyrus</td>
<td>-4 54 32</td>
<td>4.7677</td>
</tr>
<tr>
<td></td>
<td>angular gyrus</td>
<td>-40 -72 36</td>
<td>4.1218</td>
</tr>
<tr>
<td><strong>L parietal</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>posterior superior temporal sulcus</td>
<td>50 -60 28</td>
<td>4.3747</td>
</tr>
<tr>
<td><strong>R temporal</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>anterior superior frontal gyrus</td>
<td>6 54 30</td>
<td>4.295</td>
</tr>
</tbody>
</table>

Figure 2-3: Positive correlation with lexical entropy (orange clusters in the dorsal prefrontal cortex), negative correlation with averaged frequency (red clusters in the inferior frontal gyrus and the inferior temporal lobe) and positive correlation with the combined effect (yellow clusters in the inferior frontal gyrus and posterior temporal lobe) in the left hemisphere presented on a two dimensional representation of an 'inflated' cortex. The group t map was thresholded at $p < 0.001$, uncorrected, cluster threshold=20.
2.9.2 Neural correlates of syntactic entropy and averaged number of derivation steps

Peaks of positive correlation with syntactic entropy and peaks of positive correlation with syntactic complexity (averaged parser steps) \((P < 0.001\) uncorrected, cluster threshold \(=20\)) are reported in tables (2.11) and (2.12) respectively. Table (2.13) reports the peaks of the positive effect of the combined measure (syntactic entropy and complexity):

<table>
<thead>
<tr>
<th>Region</th>
<th>Location</th>
<th>MNI coordinates ((x,y,z))</th>
<th>(t) value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>L temporal</strong></td>
<td>temporal pole</td>
<td>-52 4 -26</td>
<td>7.4547</td>
</tr>
<tr>
<td></td>
<td>middle temporal gyrus</td>
<td>-62 -32 -6</td>
<td>6.13</td>
</tr>
<tr>
<td></td>
<td>inferior temporal sulcus</td>
<td>-56 -26 -8</td>
<td>5.2974</td>
</tr>
<tr>
<td></td>
<td>superior temporal sulcus</td>
<td>-52 -32 -2</td>
<td>5.0065</td>
</tr>
<tr>
<td><strong>L frontal</strong></td>
<td>IFG: triangularis part</td>
<td>-60 16 10</td>
<td>5.9256</td>
</tr>
<tr>
<td><strong>L parietal</strong></td>
<td>angular gyrus</td>
<td>-62 -56 26</td>
<td>5.2966</td>
</tr>
<tr>
<td><strong>R temporal</strong></td>
<td>temporal pole</td>
<td>48 14 -32</td>
<td>5.7645</td>
</tr>
<tr>
<td></td>
<td>superior temporal sulcus</td>
<td>60 -26 0</td>
<td>4.6949</td>
</tr>
<tr>
<td></td>
<td>middle temporal gyrus</td>
<td>54 -28 -8</td>
<td>4.646</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Region</th>
<th>Location</th>
<th>MNI coordinates ((x,y,z))</th>
<th>(t) value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>L frontal</strong></td>
<td>middle frontal gyrus</td>
<td>-50 18 38</td>
<td>6.2734</td>
</tr>
<tr>
<td></td>
<td>superior frontal gyrus</td>
<td>-24 32 48</td>
<td>4.7529</td>
</tr>
<tr>
<td></td>
<td>anterior prefrontal</td>
<td>-14 52 -6</td>
<td>4.406</td>
</tr>
<tr>
<td><strong>L parietal</strong></td>
<td>angular gyrus</td>
<td>-46 -68 40</td>
<td>7.3287</td>
</tr>
<tr>
<td><strong>R temporal</strong></td>
<td>temporal pole</td>
<td>48 16 -38</td>
<td>6.8113</td>
</tr>
<tr>
<td><strong>L cingulate</strong></td>
<td>superior parietal gyrus</td>
<td>-36 -68 48</td>
<td>3.9918</td>
</tr>
<tr>
<td><strong>R frontal</strong></td>
<td>anterior prefrontal</td>
<td>10 60 4</td>
<td>4.8552</td>
</tr>
<tr>
<td><strong>R parietal</strong></td>
<td>angular gyrus</td>
<td>52 -64 42</td>
<td>6.7421</td>
</tr>
<tr>
<td></td>
<td>precuneus</td>
<td>54 -68 34</td>
<td>4.6398</td>
</tr>
</tbody>
</table>

The syntactic factors had a bilateral effect though the activation in the left hemisphere was more pronounced (2-4.A-D). The syntactic entropy measure had a bilateral
Table 2.13: Sentence combined syntactic entropy and complexity results

<table>
<thead>
<tr>
<th>Region</th>
<th>Location</th>
<th>MNI coordinates (x,y,z)</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>R frontal</td>
<td>inferior frontal sulcus</td>
<td>32 20 28</td>
<td>5.7236</td>
</tr>
<tr>
<td></td>
<td></td>
<td>40 18 28</td>
<td>4.6777</td>
</tr>
<tr>
<td></td>
<td></td>
<td>38 24 34</td>
<td>3.7889</td>
</tr>
<tr>
<td>L temporal</td>
<td>temporal pole</td>
<td>-54 -4 -30</td>
<td>9.6486</td>
</tr>
<tr>
<td></td>
<td>middle temporal gyrus</td>
<td>-64-30 -10</td>
<td>8.6678</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-58 -32 -4</td>
<td>7.6427</td>
</tr>
<tr>
<td>L frontal</td>
<td>anterior superior frontal gyrus</td>
<td>-8 54 42</td>
<td>7.036</td>
</tr>
<tr>
<td></td>
<td>anterior middle frontal gyrus</td>
<td>-20 48 34</td>
<td>6.4388</td>
</tr>
<tr>
<td></td>
<td>middle frontal gyrus</td>
<td>-32 18 48</td>
<td>6.6225</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-32 16 36</td>
<td>4.3584</td>
</tr>
<tr>
<td></td>
<td>superior frontal gyrus</td>
<td>-10 12 70</td>
<td>6.4741</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-8 22 62</td>
<td>5.7136</td>
</tr>
<tr>
<td></td>
<td>orbital gyrus</td>
<td>-16 16 60</td>
<td>5.6191</td>
</tr>
<tr>
<td></td>
<td>IFG: triangularis part</td>
<td>-60 18 8</td>
<td>5.7322</td>
</tr>
<tr>
<td></td>
<td>orbital gyrus</td>
<td>-64 18 8</td>
<td>5.7322</td>
</tr>
<tr>
<td></td>
<td>antero-ventral prefrontal</td>
<td>-32 58 0</td>
<td>4.3323</td>
</tr>
<tr>
<td></td>
<td>anterior superior frontal gyrus</td>
<td>-20 34 50</td>
<td>4.5447</td>
</tr>
<tr>
<td>L parietal</td>
<td>angular gyrus</td>
<td>-54 -56 28</td>
<td>6.8821</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-56 -58 44</td>
<td>6.4302</td>
</tr>
<tr>
<td></td>
<td>precuneus</td>
<td>-50 -62 38</td>
<td>6.3917</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-60 -64 26</td>
<td>5.9990</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-4 -54 34</td>
<td>5.6501</td>
</tr>
<tr>
<td>R temporal</td>
<td>temporal pole</td>
<td>50 6 -40</td>
<td>9.0021</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48 16 -38</td>
<td>8.3508</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48 14 -26</td>
<td>7.2083</td>
</tr>
<tr>
<td>R frontal</td>
<td>anterior superior frontal gyrus</td>
<td>6 62 4</td>
<td>6.8148</td>
</tr>
<tr>
<td>R parietal</td>
<td>angular gyrus</td>
<td>54 -64 42</td>
<td>6.0382</td>
</tr>
<tr>
<td></td>
<td></td>
<td>58 -58 24</td>
<td>5.7963</td>
</tr>
<tr>
<td></td>
<td></td>
<td>52 -62 34</td>
<td>5.3653</td>
</tr>
</tbody>
</table>

effect in the temporal pole and the mid portion of the middle temporal gyrus and a left lateralized effect in the inferior frontal gyrus and the angular gyrus. The syntactic complexity measure (averaged number of derivation steps) had a bilateral effect in the angular gyrus and the precuneus and a left lateralized effect in the dorsal prefrontal cortex. The combined measure had a bilateral distributed effect in the temporal lobe, extending from the posterior superior temporal lobe (TPJ) and extending all the way to the temporal pole. A bilateral effect was also observed in the medial prefrontal cortex, the precuneus and the angular gyrus. A left lateralized effect was observed in the left inferior and dorsal prefrontal cortex.
Figure 2-4: A: Positive correlation with syntactic complexity (orange clusters in the middle frontal gyrus and angular gyrus), syntactic entropy (red clusters in the inferior frontal gyrus, temporal pole and mid temporal lobe) and the combined effect (yellow) in the left hemisphere. B: Medial view of positive correlation with syntactic complexity (orange clusters in the precuneus and anterior prefrontal cortex) and the combined effect (yellow clusters in the precuneus and superior and anterior prefrontal cortex) in the left hemisphere. C: Positive correlation with syntactic complexity (orange clusters in temporal pole and angular gyrus), syntactic entropy (red clusters in the temporal pole and mid temporal lobe) and the combined effect (yellow) in the right hemisphere. D: Medial view of positive correlation with syntactic complexity (orange clusters in the precuneus and anterior prefrontal cortex) and the combined effect (yellow clusters in the precuneus and anterior prefrontal cortex) in the right hemisphere. All data presented on a two dimensional representation of an 'inflated' cortex. The group t map was thresholded at $p < 0.001$, uncorrected, cluster threshold=20.
2.10 Discussion: Sentential measures

2.10.1 Lexical entropy and averaged frequency

The most striking observations regarding the neural correlates of the lexical measures at the sentence level is the near absence of any activation in the right hemisphere, the lack of anterior or mid temporal activation and the absence of unique contribution of lexical entropy outside BA 8 and the fusiform.

Averaged lexical frequency had a unique contribution throughout the inferior frontal gyrus (Broca’s area). Activation in Broca’s area for the analogous measures in the word level model was ambiguous between the two measures. This result is in line with the view of Broca’s area as mediating controlled semantic retrieval (Wagner et al., 2001). It is also consistent with the general finding in the single word processing literature regarding the effect of frequency. We have suggested, following Lau et al. (submitted), that Broca’s area is particularly engaged in cases of lexico-semantic conflict during integration of the retrieved lexical item in context (post lexical access competition). Under this view, the fact that frequency is a better estimate of competition than surprisal (since the former, but no the latter had an effect in Broca’s area) is expected since low frequency words are more likely to have higher frequency competitors compared to high frequency words. However, this interpretation will be tested with a more direct modeling of competition in future analyses.

In the Word model we observed an effect of frequency in the MTG not observed in the Sentence model. This effect of frequency was interpreted (contrary to the effect seen in Broca’s region) as an effect of the pre-activation level of the lexical item on early lexical access. Review of the literature suggests that an early effect of frequency is detectable but is less robust than the later (presumably post lexical access) effect (cf. Hauk et al. 2008 for discussion). The proposal that this early ‘hard to detect’ effect is localized to the MTG is in line with the fact that only one paper out of 6 recent papers using fMRI to investigate the effect of frequency reports activation in this region (Kronbichler et al., 2004), while all 6 report IFC activation, which we assume supports the late, post lexical access process. The fact that this effect did
not 'survive' in the Sentence model here provided further support to this view since the sentence measure can be thought of as entropy or uncertainty, which is associated with prediction. Since we propose that the frequency effect observed in the MTG is related to familiarity rather than prediction, the lack of uncertainty effect is expected.

2.10.2 Syntactic entropy and syntactic complexity

Contrary to the lexical measures, the sentential effect of the averaged syntactic measures had a larger spatial extent compared to the word level model. In particular, the combined measure (entropy and complexity) had a significant effect in the inferior frontal gyrus, both ventrally (the orbital gyrus) and more laterally (BA 45), and the superior frontal gyrus. Syntactic entropy had a unique significant effect in the left temporal pole, mid temporal regions and the inferior frontal gyrus while the related word level syntactic surprisal had no significant effect in the left hemisphere. Syntactic complexity, which had only an effect in the angular gyrus in the Word model made an additional unique contribution to activation in the left dorsal prefrontal cortex in the Sentence model.

Syntactic entropy's unique contribution to activation in the temporal pole is in line with the proposal that the ELAN is generated in that region and reflects syntactic predictability (2.5.3). Interestingly, contrary to the ERP literature, the effect here was observed bilaterally.

The fact that the mid portion of the middle temporal gyrus, implicated by the meta-analysis in Indefrey and Levelt (2004) in the storing of lemma information, exhibits unique effects of syntactic surprisal supports a probabilistic (predictive) activation of lemma information during sentence processing (cf. Van Berkum et al. 2005). This activation might also contribute to the ELAN effect.

Syntactic complexity activated both the angular gyrus and the precuneus in the Sentence model, as in the Word model, in line with our view of these regions, expressed earlier, as directly implicated by syntactic processing. The Sentence model revealed syntactic complexity effects in the dorsal part of the middle frontal gyrus. The activation cluster observed here is very close to the f2p peak in the meta-analysis.
reported in Vigneau et al. (2006). In line with the results here, these authors describe F2p as the center of a ‘purely’ syntactic processing region and make it a part of the sentence/syntax working memory loop (together with the left TPJ).

The effect of the combined measure in the orbital region of the frontal cortex (BA 47) is in line with Dronkers et al.’s (2004) finding that lesions to this region produce sentence level deficits. Though it has been traditionally associated with reward related activation (Rolls, 2000) both in monkeys and humans, recent imaging literature has implicated this region in a variety of other tasks and in particular memory (Wagner et al., 2001). In fact, recent monkey lesion study by Browning and Gaffan (2008) suggests that the frontal involvement in reward structure might be in the representation of complex temporal sequences rather than direct representation of the reward itself. These findings, coupled with the known connectivity between the orbital region and the temporal pole and the precuneus, and with the data presented here proposes that BA 47 plays an important role in the representation and manipulation of syntactic structure.

### 2.11 General discussion

From a cognitive perspective, the results reported in this chapter bear directly on a number of the issues raised in the introduction. First, we have identified independent syntactic and lexical predictive networks. At the word level, lexical prediction seems to play a more important role, compared to syntactic prediction. This situation is reversed in the sentence level. This difference in granularity is important for the conceptualization of the relation between probabilistic syntactic and lexical knowledge and processing, in an eventual parsing model.

We have identified a role both for prediction and familiarity in pre-activation. The specific distribution of the two effects suggests that early phonological and lexical processing is facilitated by familiarity but not contextual information. This result supports the view maintained by bottom-up models of word recognition and argues against early top-down control of these processes. Contextual predictability appears
to have its effect during integration of the lexical information into the sentence frame. Importantly, these are inferences based on localization of function. Data from MEG might help to (dis)confirm these inferences with more fine grained temporal analysis.

Despite the correlation between syntactic complexity and surprisal we have found, in particular in the Sentence model, that both play a role in comprehension. It is important to remember that the syntactic complexity measures used here represent the specific assumptions made in the Roark parser. More work is required to find out whether these results are robust for different assumptions regarding the syntactic structure or the parsing architecture.

Surprisal and entropy are not easy to separate using fMRI in a task like the one here since the slow nature of the BOLD response does not allow direct evaluation of each event’s surprisal effect but rather provides us with a sort of running average measure. Despite this limitation we have identified interesting differences between the two related measures. In particular, syntactic entropy, or the state of uncertainty resulting from the overall surprising nature of the syntactic structure, seems to play a more important role than lexical entropy. The lexical effect that seems to be preserved at the sentence level is the overall extent of post-lexical competition (reflected in Broca’s region).

From a neurological perspective, our results confirm and nuance the role of known language regions and suggest a role for regions previously not considered to play a role in language comprehension.

Overall our results are consistent with current assumptions regarding the mapping of linguistic processing onto the temporal lobe. One novel observation in this region was the apparent role of the Word Form Area in the processing of auditory lexical information. In addition, our results support the view that the temporal pole plays a role in syntactic predictions.

In the frontal lobe our results support the more recent view coming from lesion studies that Broca’s area itself does not play a direct role in syntactic processing and that in fact it is adjacent regions (orbital and middle frontal gyrus) that support syntactic processing. Instead, and in line with work in the memory and priming
literatures, Broca's area seems to play a role in lexical competition.

Finally, the results here argue for a central role in processing for two regions in the parietal lobe, namely the angular gyrus and the precuneus. These regions have not been a central part of the classic language network but recent evidence reviewed here and our results argue differently. In particular, we suggested that the angular gyrus supports the lexical-syntactic interface and that the precuneus supports the representation of hierarchical structure. Results in the following chapters will provide further evidence for this position.
Chapter 3

Ambiguity

*Time flies like an arrow; fruit flies like a banana* (Groucho Marx)

3.1 Structural ambiguity in sentence processing

Structural ambiguity, or the fact that the same word string can correspond to more than one syntactic structure, is pervasive in natural language. It has been one of the fundamental observations that motivated researchers such as Chomsky (1975a) to investigate grammatical formalisms more powerful than finite state automata. However, as has been observed by both psycholinguists and NLP engineers, this property provides a formidable challenge to the parser. In this chapter we will attempt to evaluate two possible effects of ambiguity on neural activity during sentence parsing:

(14) What is the effect, if any, of uncertainty generated by (local) ambiguity on brain activation?

(15) What is the effect of disambiguation towards a dis-preferred parse on brain activation?

All psycholinguistic and neuro-linguistic work to data on the topic of structural ambiguity has made use of a factorial design with binary contrasts (e.g. ambiguous versus not-ambiguous, disambiguation towards a preferred parse versus disambiguation towards a dis-preferred parse). The advantage of this approach is that it allows
us to test specific aspects of the parsing process while leaving other aspects constant across conditions. In this way, the experimenter is not required to provide a full parsing algorithm and can stay agnostic to the aspects of ambiguity processing not tested directly.

In the experimental design used here, this approach is not ideal. One could select a limited subset of the narratives that exemplifies a specific aspect of ambiguity, but by doing so, one will lose the full power offered by modeling the entire time course of comprehension.

In order to make full use of the experimental paradigm, we looked to quantify the extent of ambiguity and reanalysis at every word and to replace the binary contrast with a continuous measure. To this end, we made use of a computationally implemented probabilistic parser (Roark, 2001). The ambiguity measures were approximated using the internal state of the parser at every word. Uncertainty regarding the correct parse (14) was approximated using the entropy measure over the set of partial analyses considered by the parser at every word (the beam).

Beam entropy is related but distinct from the notions of surprisal and entropy discussed in chapter 2. There, entropy or surprisal quantified uncertainty regarding the identity of future input. Beam entropy, on the other hand, quantifies uncertainty regarding the correct analysis of the already received input (or prefix).

Disambiguation towards a dis-preferred parse was quantified as the extent of re-ranking of the top analysis within the beam, upon the integration of a new word. Re-ranking can be thought of as the parallel parsing equivalent (see below) of reanalysis in a serial parser. As with the uncertainty measure, our re-ranking measure is not binary but can take any value from 0 (no re-ranking) to 1 (when the former top analysis has fallen off the beam).

The advantage of the approach here is that we now have continuous ambiguity measures for each word. However, this approach requires the use of a fully implemented parser and so forces us to be explicit about all aspects of ambiguity processing and not only the two aspects in (14,15). In particular, we need to commit to specific answers to the following questions:
Is processing serial or parallel? (16)

Are decisions made deterministically or probabilistically? (17)

What sources of information are available or used? (18)

Does the parser attempt to resolve ambiguities as soon as possible or is indeterminacy tolerated? (19)

Roark's parser (used here) is parallel (16) and probabilistic (17). It uses statistical syntactic and lexical information (18) and tolerates indeterminacy (19). As will become evident in the discussion below, some of these features are directly relevant to the questions asked here while other are less so. We will first elaborate on the different issues and then present data from the literature that bears on them.

3.1.1 Parallel versus serial parsing

An incremental parser is considered to operate in a parallel manner if, for a given prefix, the parser can compute, maintain, and update with incoming information, more than a single structural description of that prefix (MacDonald, 1994; Gibson and Pearlmutter, 2000; Clifton and Staub, 2008). An alternative, serial, parser maintains a single structural description of the prefix.

3.1.2 Deterministic versus probabilistic parsing

A parser is deterministic or a fixed-choice parser (in the terminology of Van Gompel et al. 2005) if given a specific choice point (onset of ambiguity) the parser always follows the same course. If the choice can vary, the parser is considered probabilistic. While this distinction is usually addressed explicitly in serial frameworks, it is also relevant in a parallel architecture. A deterministic parallel parser would pursue all analyses to the same extent. A probabilistic or ranked parser (as the one used here) ranks the different alternatives under consideration and could also eliminate some alternatives with very low probability from the alternative set (Gibson, 1991).

A fixed-choice serial parser is, in a sense, blind to the ambiguity since there is no real choice. The parser simply follows a deterministic algorithm (as in Frazier 1978's...
A probabilistic serial parser is not restricted in that way. At the choice point it might follow any one of the potential alternatives. The stochastic choice would be governed by the probability distribution over these alternatives (Traxler et al., 1998).

3.1.3 Sources of information in ambiguity resolution

The debate regarding the information used by the parser in making a choice at the onset of an ambiguity (if serial) or in the ordering of the alternative analyses (if parallel) has traditionally opposed a syntax first/only view (Frazier, 1978) to a view where multiple sources of information are accessed simultaneously as in MacDonald’s (1994) constraint satisfaction architecture (cf. Elman et al. 2005; Kennison 2001 for recent contributions to this debate). Although we are still far from a consensus, the view that the human parser makes use of multiple sources of information (e.g. syntactic, lexical, pragmatic) during parsing (and in particular ambiguity resolution) has come to prevail in the last decade or so (MacDonald et al., 2006). The debate has largely shifted to the relative importance of the different sources of information and to a certain degree the time course of their integration (cf. Novick et al. 2008 for a recent discussion).

It is important to keep in mind that the question of sources of information is independent of whether the parser is probabilistic or not. In other words, two probabilistic parsers can differ with respect to which information is used to shape the distribution of probability over the alternatives. One issue that is still debated is whether statistical information plays a role in shaping that distribution. For example, the parser might make use of verb specific information such as thematic roles or possible subcategorization frames but might not use statistical knowledge regarding the use of that verb.

Though parsing accounts often conflate lexical guidance with statistical guidance these are logically independent. Jennings et al. (1997) showed that on-line naming latency (in an ambiguous context) correlated with the frame bias of the specific verb. On the other hand, Kennison (2001) and Pickering et al. (2000) presented evidence
against frequency or statistically guided parsing. It appears that the question of the relation between statistical distribution and linguistic behavior in the context of ambiguity is still unresolved.

A more fundamental question is how to interpret a correlation between linguistic behavior and corpus frequency. Is the subject attempting to replicate the corpus statistics or do the corpus statistics simply reflect a distribution that is determined by other factors which need to be uncovered?. This question goes beyond the scope of our present inquiry. The parser used here weighs different options according to their probability given the parser's grammar and training corpus. In essence, this approach makes use of both grammatical considerations and statistical ‘world knowledge’ regarding the distribution of sentences in a corpus.

Though the parser here makes use of corpus frequencies, we prefer to stay agnostic regarding the question of whether this is true of the human parser as well. We propose, instead, to take the statistical component as a place holder for the correct source of probability distribution. Statistical approximation of the training corpus allows our parser to mimic the decisions made by the human parser even if the underlying architecture is different.

### 3.1.4 Tolerating indeterminacy

It is an empirical question whether the parser (either serial or parallel and either deterministic or probabilistic) attempts to resolve the ambiguity as soon as it arises or whether ambiguity (or indeterminacy) can last until a later point. Certain serial parsers such as Frazier (1978) or Traxler et al. (1998), make an early choice¹. This is also the case in a number of prominent parallel parsing proposals such as McRae et al. (1998); Tabor and Tanenhaus (1999) where multiple analyses are generated at every choice point but processing continues only after one has been selected via competition (cf. Elman et al. 2005; Van Gompel et al. 2005; Green and Mitchell 2006 for discussion).

---

¹Putting aside the issue of possible lookahead
This early commitment provides a straightforward explanation for the garden path effect (as in the auxiliary/main verb ambiguity in 20). Once an early commitment is falsified, a reanalysis is required. One of the problems with this account of garden path is that not all cases of ambiguity lead to an experience of a garden path (as the direct object/sentential complement subject ambiguity in 21):

(20) The dog that I had really loved bones.

(21) The historian read the book had been destroyed in the flood

Since both cases require a reanalysis in an early commitment model, such models must specify why certain reanalyses are easier than others (cf. Pritchett 1992 for one such proposal).

Cases such as (21) motivated a different class of models that can accommodate certain degree of underspecification or indeterminacy. A serial parser can build underspecified arcs (as in Weinberg’s 1993 proposal that the parser initially specifies dominance rather than immediate dominance relation between constituents) or underspecify the scope of a modifier as in Frazier and Clifton (1996).

A parallel parser, as the one used here, can allow for indeterminacy by dropping the requirement that a single analysis be selected at every stage. Instead, a plurality of analyses can be maintained until either new information or a top-down directive eliminate all but the top most analysis. In fact, from the perspective of a framework that allows for underspecification it is possible that, at least in certain contexts, listeners never disambiguate a globally ambiguous sentence. We shall return to this point below.

We can now observe that from the perspective of a deterministic serial parser there is no uncertainty in the sense of (14). In this respect, (14) is made meaningful by our choice to make use of a probabilistic parser. However, in an early commitment parser uncertainty is at play only at decision points (onsets of ambiguity). Our choice of a ‘tolerant’ parser allows for uncertainty to extend throughout the ambiguous region. If we had used an early commitment parser instead, the predictions regarding uncertainty would have potentially been very different.
In an early commitment parser, disambiguation requires reanalysis of the prefix (since the alternative analyses have been discarded). The factors contributing to cognitive load of reanalysis predicted by such parsers (cf. Van Dyke and Lewis 2003 for one specific proposal) are potentially very different from the factors that affect the extent of re-ranking in a parallel tolerant parser as the one used here. As a consequence, our choice of parser has a direct effect on the measure we computed to answer the question in (15).

In summary, it is evident that the choice of parser is not innocent and could have an important effect on the predictions put to test. Ideally, we would like to test predictions made by different parsing architecture and eventually even be able to make use of a design similar to the one here to chose among competing models. This project is beyond the scope of this chapter. Instead, in the next section, we will present data from the literature that bears on the issues raised above. We will conclude that the critical aspects of the parser architecture used here are, to a large degree, supported by existing data. We will then proceed to answer the questions in (14,15), using the measures produced by the Roark parser.

3.2 Evidence for a probabilistic parallel tolerant parser

In this section we will briefly evaluate existing evidence supporting or undermining the architecture adopted here.

3.2.1 Evidence that the parser computes multiple analyses

Certain manipulations suggest that the human parser computes a less preferred interpretation. For example, Pearlmutter and Mendelsohn (ms) have demonstrated that the plausibility of the less preferred parse affects reading time of locally ambiguous sentences. In both (22) and (23) the embedded clause is initially ambiguous between a relative clause and a sentential complement (since the noun report allows, but does
not require, a complement).

(22) The report that the dictator described the country was clearly false.

(23) The report that the dictator bombed the country was clearly false.

In (22) the relative clause reading of the prefix *The report that the dictator described* ('the report described by the dictator') is plausible while the relative clause analysis of the prefix *the report that the dictator bombed* in (23) ('the report bombed by the dictator') is not very plausible. In contrast both sentential complement readings are equally plausible.

In earlier work, the authors have established that the sentential complement interpretation in this type of constructions are the preferred analysis. For an early commitment serial parser, this observed preference means that at the onset of the ambiguity (*that*) the parser follows only the sentential complement analysis. Since this analysis is never disconfirmed in either condition, such a parser never generates the alternative relative clause analysis. As a consequence, the difference in relative clause plausability between (22) and (23) should not have any effect on comprehension.

Contrary to the expectations of a serial parser, subjects read slower (22). This result strongly suggests that (at least at one point) the relative clause analysis is computed\(^2\). Other research has provided off line evidence for the computation of less preferred analysis such as syntactic priming (Van Gompel et al., 2006) and semantic interference (Sturt, 2007).

\(^2\)Lewis (2000), in defense of the serial perspective, suggests that these data could be accommodated by a probabilistic serial parser since such parser would pursue the relative clause interpretation at least some of the time and so, on average, the implausible sentence would be read slower. This proposal makes the prediction that we should find a trial by trial positive correlation between reading times on the ambiguous region and the disambiguation region, since the trials that exhibit plausibility effects are predicted to involve a serial choice of a relative clause analysis and so would require reanalysis.
3.2.2 Lack of competition effects during the ambiguous region

Early choice parallel models predict competition effects at certain choice points during parsing. Implemented parallel models such as Tabor et al. (1997); McRae et al. (1998); Elman et al. (2005) make this cost explicit, modeling it as the number of steps (or network iterations) required for the parser to select an analysis in the presence of ambiguity. When the two (or more) possible analyses have equal weight (or probability), arriving at a decision is predicted to take longer and hence increase cognitive load compared to a non ambiguous control, or ambiguous regions with one highly preferred analysis (though see Green and Mitchell 2006 for some nuances).

Despite the apparent straightforward nature of the prediction, the experimental results have been contradictory and inconclusive (cf. Clifton and Staub 2008 for a recent review). While competition effects for lexical ambiguity are well attested behaviorally, such effects in the case of structural ambiguity are not so. According to recent reviews (Clifton and Staub, 2008; Clifton et al., 2007), there is not a single unambiguous piece of evidence in the eye tracking literature for increased processing cost during a syntactically ambiguous region. This finding is particularly striking given the demonstrated effect of lexical ambiguity (Sereno et al., 2006) as well as syntactic disambiguation (Ferreira and Clifton, 1986) on eye movement during reading.

However, brain imaging and electrophysiology do suggest a cost for the presence of the ambiguity (or the multiple analyses). Frisch et al. (2002) used the fact that the feminine determiner in German is ambiguous between an accusative and nominative case while the masculine is not to investigate the effect of ambiguity on the processing of sentence initial noun-noun sequences (subject-object or object-subject). They com-

---

3 Clifton et al. (2007) conducted a meta-analysis of about 100 eyetracking papers looking at syntactic effects on eye movement measures. Though few of them did report increased reading time during the ambiguous region, in all but one (Ni et al., 1996) the effect of this factor was confounded with the presence of a functional word (usually that) preceding the unambiguous control. Ni and colleagues used the tense of the verb to disambiguate avoiding this confound. However, in the material of this last study lexical items varied across conditions, making any strong conclusion problematic.
pared ERP response to sentences containing an ambiguous feminine noun phrase in initial position (disambiguated by the following, unambiguous, noun phrase) to sentences with unambiguous accusative and nominative masculine noun phrases. The authors found an increased positivity in the ambiguous condition over both medial and lateral electrodes around 600ms after the onset of the presentation of the noun. While this difference is unarguably due to the ambiguity, it is possible to interpret it as a lexical, rather than structural, ambiguity of the determiner.

Mason et al. (2003) (experiment 2) used fMRI to compare reading of sentences with local reduced relative ambiguity with reading of unambiguous controls. The ambiguous sentences were disambiguated either towards the 'preferred' main verb reading or towards the dispreferred reduced relative clause reading. The authors found that disambiguation towards the reduced relative produced the greatest signal change in their ROI's (in the left inferior frontal gyrus and in the temporal gyrus). Critically, however, the sentences disambiguated towards the preferred interpretation also produced larger activation compared to the unambiguous control (cf. Novais-Santos et al. 2007 for similar results). These results suggest that the ambiguity itself produces increased neural activation, independently of reanalysis. However, this inference assumes a model where only disambiguation towards a less preferred parse is costly.

Finally, two fMRI studies in the literature manipulated the length of a structurally ambiguous region (Fiebach et al., 2004; Novais-Santos et al., 2007). The first manipulated the point at which the grammatical role of an ambiguous sentence initial NP in German was disambiguated (early versus late in the sentence). The second study added optional material between the onset of a direct object/embedded subject ambiguity and the point of disambiguation. Both studies found an effect of ambiguity length. Fiebach et al. (2004) found a positive effect of length in the superior portion of BA 44 (pars Opercularis in Broca’s region), inferior parietal and the right basal ganglia. Novais-Santos et al. (2007) found a length effect in the left inferior parietal

4And possibly the noun, depending on one's assumptions regarding the representation of gender in the German mental lexicon.
and angular gyri. As with Mason et al. (2003), the effect of length held for both the preferred and dispreferred analyses, suggesting, again, that it was not due to an increased effort during reanalysis (Van Dyke and Lewis, 2003). Novais-Santos et al. (2007) found a reanalysis effect (embedded subject versus direct object) in the dorsal prefrontal cortex while ambiguity length had its effect in the left parietal lobe.

The behavioral and neural evidence seem to be contradictory. The neural evidence suggest a cost for the ambiguity itself but behavioral measures do not reflect such a cost. However, it is possible for a cognitive process not to have a behavioral consequence and from that perspective neuroimaging measures can be more sensitive than behavioral measures. Also, two opposing effects could 'cancel out' a behavioral effect but still be distinguished neurally. In the next section we review evidence that suggests this might be the case here.

3.2.3 Ambiguity is not always resolved

In a series of papers, Pickering and colleagues (Traxler et al., 1998; Van Gompel et al., 2001, 2005) have demonstrated, using eye tracking, that in balanced ambiguity (24), reading time of the globally ambiguous sentences (24c) is shorter than either disambiguated (by the anaphor) variants (24a,24b):

(24) a. The daughter_i of the colonel_j who shot herself_i\_j on the balcony had been very depressed.

b. The daughter_i of the colonel_j who shot himself_i\_j on the balcony had been very depressed.

c. The son_i of the colonel_j who shot himself_i\_j on the balcony had been very depressed.

Van Gompel et al. (2005), replicating these results with other ambiguous configurations, added to the paradigm a non-ambiguous variant and demonstrated that the globally ambiguous condition was read as fast as a non-ambiguous control. The fact that the two disambiguated variants did not differ on any of the eye movement measures argues against a fixed-choice serial parser since such a parser would commit
to either a high or a low attachment in all variants of (24). In either (24a) or (24b) this commitment would consistently require revision at the anaphor and we would expect that variant to be read slower.

Pickering and colleagues take this fact to argue in favor of their probabilistic serial parser, the unrestricted race model, where the initial commitment is not deterministic but is determined by contextual (syntactic and semantic) constraints or cues. Since their data consisted of balanced cases of ambiguity, the probabilistic parser would chose the higher attachment 50% of the lower attachment the other 50%. This would lead to 50% cases of revision in both the disambiguated conditions, resulting in slower reading times. Since no disambiguation takes place in (24c), no matter what was the original commitment of the parser, there would be no cause for slowing up, hence the apparent advantage of the ambiguous condition.

Levy (2008) proposes an alternative explanation for the pattern in (24) based on surprisal. Specifically, Levy proposes that the anaphor (himself) in (24c) is supported by two distinct analyses (low and high attachment) while in either disambiguated version the anaphor is compatible with only one of the two analyses. Since the prefix probability is calculated as the sum of the probabilities of all analyses compatible with the prefix, the prefix probability at the anaphor in (24c) will be strictly greater than in either (24a) or (24b) and so the surprisal value of the anaphor will be lower.

It is important to note that the parser Levy has in mind (he is not explicit on this point) is of the delayed-resolution kind. In the context of 14, Levy is essentially proposing that higher entropy, associated with the absence of resolution, produces faster reading rate (or less cognitive effort; cf. Hale 2006 for a related proposal). This perspective makes an interesting prediction not shared by the serial unrestricted race model. Since no information disambiguates between the two analyses in (24c), by the end of the sentence the parser will not commit to either (as long as they are truly unbiased) or at least the less preferred interpretation should still be among the competing analyses (while in the disambiguated cases it has been clearly discarded from the beam on all accounts). In the unrestricted race model, the parser has committed
to a single analysis from the get go and so and no indeterminacy is expected.\footnote{The unrestricted race model does expect a bimodal distribution across trials in the ambiguous condition since, in the absence of biasing constraints, the parser choses stochastically one interpretation at every trial. However, per trial the parser itself has no internal ambiguity.}

Swets et al. (2008) provide data that seems to support the consequences of Levy's proposal over the predictions of the unrestricted race model. These authors replicated Traxler et al.'s (1998) experiment with the added manipulation of the comprehension question following each sentence. In the original series of papers comprehension questions were presented only intermittently and were not designed to probe the interpretation a subject selected for the ambiguous condition. Swets et al. (2008) used three question conditions. The 'occasional' condition was similar to the original design. In the 'superficial' condition a question followed every sentence but could be answered without resolution of the ambiguity (eg. Did someone shoot himself? for 24c). Finally, the 'relative clause' condition probed after each sentence the specific interpretation subjects assigned to the sentence (eg. Did the colonel shoot himself?).

The superficial and the occasional conditions replicated the reading time pattern observed by Pickering and colleagues (3-1). In these two conditions the anaphor and the following region were read faster in the globally ambiguous sentences than in either disambiguated version. Moreover, there was no difference between the two disambiguating conditions.

The relative clause condition produced very different reading patterns. First, there was a large main effect of the question type with reading times almost doubling over the target regions. Secondly, The high attachment condition (24a) was read the slowest, followed by the ambiguous version (24c). The low attachment (24b) produced the fastest reading times. These results demonstrate that the question type had an effect on subjects' reading pattern and specifically that more detailed questions produced longer reading times, suggesting a strategic choice. More importantly to the topic here, this slowing down (or more attentive reading strategy) revealed a preference for low (or local) attachment not seen in the other two conditions and eliminated the advantage of ambiguity. These results suggest that lack of disambiguation is
Figure 3-1: Reading times from Swets et al. (2008) for the postdisambiguating region (e.g., on the balcony) as a function of question type and sentence type. Error bars represent standard errors.

advantageous only if the subject is not required to evaluate the exact interpretation.

Figure 3-2: Question response times from Swets et al. (2008) as a function of question type and sentence type, for correct answers only. Error bars represent standard errors.
Analysis of the response pattern to the comprehension questions strengthens this conclusion (3-2). Answers to globally ambiguous sentences took significantly longer than to disambiguated sentences but only when the response required disambiguation (in which case subjects showed a low attachment preference).

These results are unexpected from the perspective of the serial unrestricted race model or any early resolution parser. Swets et al. (2008) propose that their data supports a late-resolution serial parser (specifically Frazier and Clifton 1996). This conclusion, however, might be problematic in view of their own data. The authors find that high attachment relative clauses are slower to read and that subjects produce significantly more errors on the following questions. The authors own interpretation is that this high error rate comes about "... perhaps because their initial N2 interpretation was not fully overridden by the ultimately correct N1 interpretation" (p. 208). But how would an initial N2 interpretation be computed if the parser specifically underspecifies the attachment site?

We would like to suggest that Swets et al. (2008) results are in fact compatible with a parallel probabilistic version of late-resolution. Together with Swets et al. (2008), and implicitly Levy, we interpret the original results from Pickering and colleagues to be a consequence of indeterminacy in the parser. In a parallel parser this indeterminacy is reflected by higher entropy over the candidate set (since the probability mass is not concentrated over one candidate). As a consequence, higher entropy should negatively correlate with cognitive load caused by 'full interpretation'. Once entropy is reduced (either by disambiguating input, by a top-down strategy or by the need to answer a question that requires a choice between the alternatives) cognitive load is expected to increase.

The results reported above are fully compatible with this suggestion. Ambiguous regions are read faster than disambiguated regions (in the absence of a top-down strategy that looks to reduce entropy) since they do not entail entropy reduction. The answer to specific questions would produce an entropy reduction only for the ambiguous sentences (since disambiguation already took place earlier on for the other two conditions) and the longer reading time is predicted.
This perspective can account for the aspect of the data that proved to be complicated for the serial underspecification view. While entropy in a probabilistic parallel parser is analogous in some respects to underspecification in a serial parser (as discussed above), probabilistic parser can encode preferences (for example, the low attachment preference clearly observed in Swets et al.'s 2008 data) despite the ambiguity, which is not possible in an underspecified serial parser. In other words, a high level of entropy does not require equidistribution of probability across all candidates (the probabilistic equivalent of underspecification), only a relatively flat (unpeaked) distribution. As a consequence, the high error rate for the dispreferred disambiguation condition can be potentially captured as well as the overall preference for low attachment interpretation seen in the responses to the ambiguous condition.

Returning to the apparent discrepancy between neural and behavioral evidence discussed above, we propose that the lack of behavioral effects of ambiguity is a consequence of the interaction of two opposing factors. Higher uncertainty (or entropy) requires more neural resources (demonstrated in the imaging literature). However, the same uncertainty also leads to suspension of interpretation which results in faster overall processing (as demonstrated by globally ambiguous sentences). As a result, the two opposing forces cancel each other in the behavioral measure. This novel perspective predicts that we will find neural activation which is negatively correlated with ambiguity as well as neural activity that is positively correlated with it. We will return to this issue in the results section.

### 3.2.4 Evidence from disambiguation cost

In principle, (late-resolution) parallel and (early-resolution) serial parsers make distinct predictions with respect to the point of disambiguation. Since parallel parsers keep all alternatives updated, there should be no cost for disambiguation. A serial parser, on the other hand, having abandoned (or never have computed) the alternative parse, would be required to perform a (costly) reanalysis of the prefix and produce a new structural analysis compatible with the new information.

It is by now well established that disambiguation towards a less likely interpreta-
(25) The psychology students found the book in the bookstore.

(26) The psychology students found the book was written poorly.

Hare et al. (2003) demonstrated, for example, that in a context supporting a direct object interpretation of *the book* in (26), subjects slow down when reading the continuation *was written*, which disambiguated the parse towards the embedded sentence analysis. This body of evidence speaks against a version of a parallel parser where all alternatives are weighted equally.

However, most parallel parsers take a probabilistic view of the alternatives and so could also predict a disambiguation effect in cases such as (25) that would result from *re-ranking* of alternatives (or the suppression of a high probability analysis in favor of a low probability alternative). This is not to say, though, that data from disambiguation could not be used to tease apart the two architectures. Gibson and Pearlmutter (2000) point out that an early-resolution serial parser predicts a bimodal/dichotomic distribution of a given dependent variable at the point of disambiguation. Either the parser had committed earlier to the correct analysis supported by the disambiguating information or a complete reanalysis is required. A probabilistic parallel parser would presumably exhibit a more graded response which would be modulated by the extent of redistribution of the probability mass over the alternatives. It is has not been easy to test these predictions since most experiments have used a two-level factorial design (preferred versus not preferred disambiguation) rather than a parametric design (Lewis, 2000).

A different way of thinking about Gibson and Pearlmutter’s (2000) point is that a parallel parser could accommodate an effect of a dispreferred analysis at the point of disambiguation. Pickering et al. (2000) compared reading patterns (using eyetracking) across pairs of sentences as in (27):

(27) a. The young athlete realized her potential one day might make her a world-class sprinter.
b. The young athlete realized her exercises one day might make her a world-class sprinter.

The authors used a number of norming tests to establish that the verb realize (both in isolation and this specific context) is associated with a strong preference for sentential complement completion over a direct object completion (as in The young athlete realized her potential). A serial lexical-guided parser should commit initially to a sentential complement analysis and so the implausibility of the dispreferred direct object reading should not affect either the reading time during the ambiguous region or disambiguation.

Contrary to this prediction, the implausible object condition (27b) produced more looking time during the ambiguous region (her exercise one day) than the plausible condition (27b). Moreover, the disambiguating region (might make) was read faster in the implausible object condition. These results suggest that readers do pursue a direct object analysis even in a context with strong bias towards the alternative analysis, and moreover, that information 'collected' during the ambiguous region (the plausibility of the direct object interpretation) can change the effect at the disambiguating region.

This pattern is expected from the perspective of a late-resolution parser. At the verb, both the direct object and sentential complement analyses are pursued. The slowing down at the NP in the implausible condition can be a result of re-ranking if the direct object was the preferred analysis or a surprisal effect as proposed by Levy since the NP in (27a) is compatible with the two analyses, while the NP in (26) is compatible only with one of the two, producing a larger drop in prefix probability. The speeding up at the embedded verb is expected as well since (by either an earlier re-ranking or surprisal) the direct object analysis in the implausible condition has been discarded (or demoted).

---

6 The term is borrowed from the authors and describes any parser which uses lexical frequency or predictability to commit to a specific single interpretation at the onset of the ambiguity. The authors, in fact, in arguing against this family of parsers, posit as the alternative a deterministic serial parser that always commits to a direct object analysis. It seems to us that the data are also compatible with lexically guided parallel parser.
3.2.5 A note on modality

One important caveat in the context of the current experiment regards modality. All the work cited above, and in fact most work on syntactic ambiguity, has used visually presented material (with the exception of the visual world paradigm, Tanenhaus et al. 1995). Our own parser estimates do not incorporate prosodic information or ambiguity due to the auditory modality. Our data, however, come from an auditory paradigm. It is possible that the two modalities differ in a way that makes this idealization too distant from reality. To the extent that our results are interpretable, this is probably not the case, but we are currently at the first stages of an attempt to produce entropy measures directly from the auditory wave form\(^7\).

3.2.6 Summary

We believe that the current inventory of evidence regarding the processing of potential ambiguities supports (or at least is compatible with) the basic assumptions of the architecture adopted here. Comprehenders pursue multiple analyses which are differentially weighted (or ranked). This ranking can be modified during an ambiguous region. Finally, at least in certain cases of ambiguity (and task demands) the parser can adopt a late-resolution strategy, not necessarily converging on a single analysis. The available behavioral and imaging data suggest that the presence of multiple alternative analyses could have both a facilitatory effect, arguably because the parse remains underspecified; and cost, possibly due to increase in memory or attentional demands. In effect, the null effects of ambiguity in behavioral measures could be a result of these two opposite effects ‘canceling out’.

In the context of the parsing architecture adopted here, the two questions that we set up to investigate (14,15) can be reformulated in the following way:

(28) What is the effect, if any, of the entropy over the beam on brain activation?

\(^7\)Novick et al.’s 2008 work is also instructive in this respect. The authors tested the same subjects on ambiguity resolution in both modalities. They found a correlation between an individual’s reliance on lexical information across modalities.
What is the effect of re-ranking on brain activation?

The first question regards the alternative stack but instead of a binary ambiguous/unambiguous comparison we suggest to view ambiguity as a graded measure and propose to quantify it as the uncertainty (entropy) at every point regarding the correct analysis of the prefix. Since we suspect ambiguity might have both positive and negative effects, the first question purposefully avoids specifying a directionality. By providing us many more simultaneous, spatially distinct, measurements, brain imaging could potentially allow us to tease apart the two effects.

The second question addresses an issue which is hard to assess in the usual two levels factorial design (see above). Ranked parallel processing suggests that not all cases of reanalysis (or re-ranking) are born equal. In particular it seems reasonable to assume that the cost of re-ranking will be positively correlated with how probable the now demoted analysis was before re-ranking and to what extent has it been disconfirmed. In the next section we will present the formal details of the calculation of these two predictions.

3.3 The current experiment

Instead of exploring a single region or a specific construction, we were interested in coming up with a quantitative measure of uncertainty and re-ranking for every word that would allow us to make full use of the power of our experimental design. In order to achieve this goal we used the internal states of Roark’s parser to produce an ambiguity measure (3.3.1) and a re-ranking or a disambiguation measure (3.3.2).

3.3.1 The ambiguity measure: beam entropy

We quantified ambiguity as the entropy, at each input, over the set of alternatives. Informally, entropy reflects uncertainty with full entropy 1 reflecting complete un-
certainty and 0 total certainty. Alternatively, it is possible to think of entropy as a measure of the peak-ness of the probability distribution over the alternatives. A very peaked distribution produces low entropy while a flat distribution would lead to higher entropy. More formally:

- Each analysis $A$ in the candidate set $C_{w_1...w_n}$ associated with a prefix $w_1 \ldots w_n$ is a sequence of production rules $d_1\ldots d_m$, each associated with a specific probability $(\rho(d))$. The probability of the analysis $A$ is defined as:

$$\rho(A) = \prod_{i=1}^{m} \rho(d_i)$$  \hspace{1cm} (3.1)

- The entropy $H$ over the candidate set $C$ is defined as:

$$H(C) = - \sum_{A \in C} \rho(A) \log(A)$$  \hspace{1cm} (3.2)

It is important to distinguish this entropy measure from the entropy discussed by Hale (2006). Hale proposes to calculate the entropy over all possible continuations of a given prefix. Here we quantify the entropy over the possible analyses of the already existing input (the prefix itself). While these quantities could be related, they can also easily diverge\(^9\).

3.3.2 Disambiguation towards a less preferred parse: Re-ranking

For each word in the sentence, reanalysis cost is greater than 0 if the top analysis (with the highest probability) was not a continuation of the top analysis at the preceding word. re-ranking cost was quantified by calculating the extent of the drop in probability of the previously leading analysis:

\(^9\)At the end of a globally ambiguous sentence, entropy over the continuation is 0, while the entropy over the candidate set can be high. Inversely, following an unambiguous NP+relative pronoun there is little doubt that the correct analysis of the current prefix should contain an S' node but the entropy over possible continuations (type of relative clause, depth of embedding) can be great.
• For a prefix \( w_1 \ldots w_n \) the \textit{top analysis} \( t_n \) is defined as the candidate analysis with the highest probability in the candidate set \( C_{w_1 \ldots w_n} \).

• The set \( T_n \in C_{w_1 \ldots w_{n+1}} \) is defined as the (possibly empty) set of analyses in \( C_{w_1 \ldots w_{n+1}} \) which are continuations of \( t_n \).

• For a given grammar \( G \), an analysis \( A \) of prefix \( w_1 \ldots w_n \) is said to be a continuation of an analysis \( B \) of the prefix \( w_1 \ldots w_{n-1} \) iff \( A \) can derive \( B \) in \( G \) \( (A \Rightarrow G B)^{10} \).

• The disambiguation value \( D \) at \( w_{n+1} \) is 0 if \( t_{n+1} \in T_n \). That is, if the top analysis at \( w_{n+1} \) is a continuation of the top analysis at \( w_n \).

• Otherwise \( D \) is equal to 1 - the ratio between the sum of probabilities of the continuations and the probability of the previously top analysis:

\[
D(w_{n+1}) = 1 - \sum_{A \in T_n} \rho(A) / \rho(t_n)
\]  

(3.3)

Notice that \( D(w_n)=1 \) (maximal reanalysis cost) if the top analysis of \( w_{n-1} \) has no continuations in the candidate set of \( w_n \). Using (3.3) to model disambiguation cost or difficulty is an attempt to capture the intuition in the literature that disambiguation difficulty hinges on the strength of one's commitment to a particular parse and the extent to which the addition of the new word changes this commitment. We could not find a quantitative formulation of this intuition (or at least one that could be extended to our formalism) and so came up with the heuristic quantification 3.3. We are well aware that there are many other possible ways to express this intuition but will be leaving this issue for further research. It is also important to note that we do not think 3.3 (or an alternative formulation) is the necessarily the only source for varying difficulty during disambiguation (cf. Van Dyke and Lewis 2003).

---

\(^{10}\text{cf for discussion of the notion derive.}\)
3.4 Results

3.4.1 Neural correlates of beam entropy

Peaks of positive and negative correlation with beam entropy ($P < 0.001$ uncorrected, cluster threshold $=20$) are reported in table (??). The ambiguity measure had a

<table>
<thead>
<tr>
<th>Region</th>
<th>Location</th>
<th>MNI coordinates (x,y,z)</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>L parietal</td>
<td>intraparietal sulcus</td>
<td>-42 -42 48</td>
<td>5.2</td>
</tr>
<tr>
<td>L frontal</td>
<td>inferior frontal sulcus</td>
<td>-46 32 20</td>
<td>4.99</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-50 40 20</td>
<td>4.04</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Region</th>
<th>Location</th>
<th>MNI coordinates (x,y,z)</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>L temporal</td>
<td>sup. temporal gyrus</td>
<td>-52 -28 2</td>
<td>5.23</td>
</tr>
<tr>
<td></td>
<td>anterior temporal</td>
<td>-48 -14 -28</td>
<td>5.18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-46 12 -38</td>
<td>5.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-52 6 -24</td>
<td>4.37</td>
</tr>
<tr>
<td></td>
<td>mid temporal gyrus</td>
<td>-54 -16 -16</td>
<td>4.7</td>
</tr>
<tr>
<td></td>
<td>sup. temporal sulcus</td>
<td>54 -28 2</td>
<td>4.66</td>
</tr>
<tr>
<td></td>
<td>planum temporale</td>
<td>-50 -22 10</td>
<td>4.56</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-58 -16 10</td>
<td>4.51</td>
</tr>
</tbody>
</table>

significant left lateralized positive effect in the intraparietal sulcus and on both banks of the inferior frontal sulcus (BA 46) (3-3). A significant negative effect of ambiguity was observed in the left hemisphere anterior portion of the temporal lobe (superior and medial temporal gyri), along the medial portions of the superior and inferior temporal sulci, and the supramarginal gyrus.
Figure 3-3: Effects of beam entropy in the left hemisphere presented on a two dimensional representation of an 'inflated' cortex. Positive effect in red (clusters in the intraparietal sulcus and dorsal prefrontal cortex). Negative effect in blue (clusters along the superior temporal lobe and temporal pole). The group t map was thresholded at $p < 0.001$, uncorrected, cluster threshold=20.
3.4.2 Neural correlates of re-ranking

Peaks of positive and negative correlation with the re-ranking measure ($P < 0.001$ uncorrected, cluster threshold $=20$) are reported in table (3.2):

<table>
<thead>
<tr>
<th>Region</th>
<th>Location</th>
<th>MNI coordinates (x,y,z)</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>L temporal</td>
<td>temporal pole</td>
<td>-50 10 -28</td>
<td>8.9594</td>
</tr>
<tr>
<td></td>
<td>mid temporal gyrus</td>
<td>-52 2 -28</td>
<td>6.8769</td>
</tr>
<tr>
<td></td>
<td>superior temporal gyrus</td>
<td>-62 -24 0</td>
<td>6.4071</td>
</tr>
<tr>
<td></td>
<td>planum temporale</td>
<td>-56 -46 16</td>
<td>4.5503</td>
</tr>
<tr>
<td></td>
<td>medial superior frontal gyrus</td>
<td>-6 68</td>
<td>5.264</td>
</tr>
<tr>
<td>R temporal</td>
<td>superior temporal sulcus</td>
<td>50 -26 -8</td>
<td>9.7669</td>
</tr>
<tr>
<td></td>
<td>superior temporal gyrus</td>
<td>62 -32 4</td>
<td>7.6685</td>
</tr>
<tr>
<td></td>
<td>planum temporale</td>
<td>66 -22 8</td>
<td>5.6973</td>
</tr>
<tr>
<td></td>
<td>precentral gyrus</td>
<td>56 4 44</td>
<td>6.6103</td>
</tr>
<tr>
<td></td>
<td>middle frontal sulcus</td>
<td>26 38 14</td>
<td>5.8156</td>
</tr>
<tr>
<td></td>
<td>opercular gyrus</td>
<td>60 12 6</td>
<td>5.4452</td>
</tr>
<tr>
<td>R frontal</td>
<td>inferior frontal sulcus</td>
<td>42 26 24</td>
<td>4.9594</td>
</tr>
<tr>
<td>L parietal</td>
<td>intraparietal sulcus</td>
<td>-36 -62 42</td>
<td>4.7106</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-46 -54 50</td>
<td>3.7866</td>
</tr>
</tbody>
</table>
Disambiguation, or extent of re-ranking positively correlated with activity in regions within the temporal lobe bilaterally with more extensive activations seen in the left hemisphere. In the left temporal lobe activation extended along the superior temporal gyrus and sulcus from the planum temporale to the temporal pole (3-4.A). In the right temporal lobe activation correlating with disambiguation was restricted to the planum temporale and the middle section of the superior temporal sulcus. In addition, correlated activation was seen along the right precentral gyrus (3-4.B). Negative correlation with disambiguation was observed only along the banks of the left posterior intraparietal sulcus.

Figure 3-4: A: Effects of re-ranking in the left hemisphere presented on a two dimensional representation of an 'inflated' cortex. Positive effect in red (clusters along the superior temporal lobe and temporal pole and inferior frontal gyrus). Negative effect in blue (posterior intraparietal sulcus). B: Effects of disambiguation in the right hemisphere presented on a two dimensional representation of an 'inflated' cortex. Positive effect in red (clusters along the superior temporal lobe and precentral gyrus). The group t map was thresholded at $p < 0.001$, uncorrected, cluster threshold=20.
3.5 Discussion

3.5.1 Ambiguity

The anterior intraparietal sulcus has been implicated by multiple studies of working memory in both visual (spatial) and verbal tasks (Cohen et al., 1997). According to a recent emerging view, the role of this region is not modality specific nor information type (serial order versus item) specific as has been proposed in the past (Brown et al. 2000; Awh et al. 1996). Instead, following Cowan's (1995) view of working memory, the intraparietal sulcus is taken to mediate the focus of attention. This region is engaged whenever a task (independent of modality, Majerus et al. 2007, or type of task, Majerus et al. 2006) requires additional working memory attention (Ravizza et al., 2004; Mayer et al., 2007). The only two fMRI studies to have manipulated length of ambiguity have both reported length related activation in (Novais-Santos et al., 2007) or near (Fiebach et al., 2004) this region.

In view of this literature, our results showing positive effect of ambiguity in the left intraparietal sulcus support a probabilistic parallel parsing architecture as they indicate that increased uncertainty (or ambiguity) correlates with increased demand on working memory or attention. A serial parser, on the other hand, does not predict such working memory cost for ambiguity. From a neuro-cognitive perspective these results seem to be captured very naturally by the focus of attention view of this region's role.

It is important to note that entropy level is not tied to the number of alternative analyses but rather to the extent of competition among them (the more flat the probability distribution, the more 'similar' in strength are the alternatives and the higher the entropy). Under this view of ambiguity cost, it would be natural to expect additional attentional resources to be engaged with increased competition. Putting aside the question of the specific mechanism by which ambiguity gives rise to activation in the intraparietal region, a probabilistic view of parsing relates ambiguity to the more general phenomenon of uncertainty. The entropy measure used here reflects directly this intuition. In this respect it is interesting to note that Volz et al. (2004);
Huettel et al. (2005) have found a correlation between intraparietal sulcus activation and decision uncertainty. The latter authors even suggest that "a primary function of posterior parietal cortex is to support the generation and modification of a set of context-appropriate response" (p.3309).

The second cluster of ambiguity correlated activation was observed in the left BA 46 region in the frontal lobe. While this region has not been observed in other syntactic ambiguity studies Shibata et al. (2007) found that this region responds to metaphoric sentences (e.g. life is a voyage) which could be interpreted (loosely) as an effect of ambiguity (between the literal and metaphoric meaning).

BA 46 has been implicated in verbal working memory (Braver et al., 2001) tasks and attention manipulations (Hopfinger et al., 2000) (cf. Badre 2008 for a review and synthesis in terms of cognitive control). Uncertainty might also increase activation in this region. Volz et al. (2004) find an effect of increased uncertainty in the right homolog of the region identified here. More generally, Koechlin and Summerfield (2007), building on their own previous work (Koechlin et al., 1999) and other work, suggest a characterization of cognitive control, which is supported by this region, in terms of informational entropy.

Recent work (Vincent et al., 2007) used correlations in spontaneous BOLD signal fluctuations in a resting state to distinguish three large scale networks. Interestingly, BA 46 and the anterior intraparietal sulcus have been identified as part of the frontoparietal control network by these authors. The results here, which found the two regions correlating with uncertainty, agree with Vincent and colleagues’ results and contribute towards a more specific characterization of the function of this network.

The temporal regions identified by the negative contrast (regions correlating negatively with the ambiguity measure) have been associated with language related functions and in particular lexical and compositional semantics (Lau et al., submitted; Hickok and Poeppel, 2007). The fact that these regions show deactivation (or less

---

142

---

11 Mason et al. 2003 have used an ROI approach and so did not consider data from this region. Fiebach et al. 2004 investigated ambiguity effect using a functional ROI and so again, did not report results from this region.
activation) in the presence of uncertainty or ambiguity could be related to the behavioral finding that comprehenders delay interpretation during ambiguity (Swets et al., 2008).

### 3.5.2 Re-ranking

The regions in the left temporal lobe that correlated positively with re-ranking have been implicated in complex auditory (planum temporale, Overath et al. 2007), linguistic auditory processing (superior temporal gyrus, Price et al. 2003; Vouloumanos et al. 2001) word level semantic processing (mid temporal) both in comprehension (Murtha et al., 1999; Noppeney et al., 2004) and production (Indefrey and Levelt, 2004; Price et al., 2005), and sentence level compositional semantics or syntax (temporal pole, Humphries et al. 2005). Activation in the right temporal lobe showed a qualitatively homologous pattern, except for the temporal pole activation.

In the left inferior cortex there was no significant activation but this is due, in part, to our thresholding. With cluster threshold set to 5, activation is revealed in the left orbital gyrus (BA 47, [-44, 26, -8]). Interestingly no activation was observed in Broca's area (BA 44,45) even when the threshold was lowered even further ($p < 0.005$).

Though the medial superior frontal gyrus has been variously implicated as a part of a distributed network of regions underlying complex linguistic processing (Michael et al., 2001; Yokoyama et al., 2006) it is not usually discussed in the context of language processing (with a few recent exceptions such as Carli et al. 2007; Alario et al. 2006; Van Heuven et al. 2008). This region has been implicated in various aspects of cognitive control such as detection of unfavorable outcomes, response errors, response conflict, and decision uncertainty (Ridderinkhof et al., 2004). In effect, one of the language studies cited above (van Heuven et al) found this region to be active in bilingual subjects in a condition that enhanced competition between the two languages. This characterization of the medial frontal gyrus provides a natural explanation for its involvement in re-ranking. Re-ranking is a form of error detection since the parser 'realizes' that the formerly preferred analysis of the sentence was probably wrong. Re-ranking could potentially also lead to a conflict between the old
top analysis and the new one.

Viewed globally, the increased activation in language processing (temporal lobe) and executive control (prefrontal cortex) regions is in agreement with the behavioral literature that has consistently reported increased reading times and error in cases of ambiguity resolution towards a less preferred analysis.

The interpretation of negative correlation with re-ranking is open to speculation. Plausibly it reflects ongoing activity that is suppressed when task demands increase. This so called ‘default mode’ activation Raichle et al. (2001) has been in the center of recent scientific research (cf. Buckner et al. 2008 for an extensive review). Interestingly, the posterior intraparietal region has been often implicated as a component of the network sustaining this default activation. Our finding, that this region decreases its activation with increased re-ranking cost is in line with this general finding.

3.6 General discussion

This chapter addressed two questions regarding the effect of structural ambiguity on brain activation:

(30) What is the effect, if any, of uncertainty generated by (local) ambiguity on brain activation?
(31) What is the effect of disambiguation towards a dis-preferred parse on brain activation?

With respect to the first question, we found that uncertainty regarding the correct analysis of the input string produces increased activation in the fronto-parietal control network and decreased activation in language processing regions within the temporal lobe. This bi-directional effect can explain the apparent lack of impact of structural uncertainty on behavioral measures as a consequence of the two opposing forces ‘canceling out’.

The fact that regions showing positive correlation with uncertainty are known to participate in working memory and attention processes supports a parallel parsing
view where uncertainty creates a greater memory load (caused by the need to maintain multiple analyses). However, further research, including direct comparison with the predictions of a probabilistic serial parser, is required to establish this conclusion more firmly.

The negative correlation in language processing regions with uncertainty fits well with proposals in the literature that, faced with ambiguity, the parser can delay interpretation. This result is not expected under early decision assumptions and hence supports an architecture that tolerates indeterminacy.

With respect to the second question, our results are in line with previous research regarding disambiguation, showing increased activation in the language network due to re-ranking of a previously preferred analysis. However, unlike previous research, we have demonstrated a graded response to re-ranking that correlates with the extent of re-ranking. Moreover, by concurrently modeling re-ranking and ambiguity this is the first study that has actually attempted to tease apart the differential effect of these two dimensions of ambiguity processing.
Theory of Mind

Theory of mind (TOM), as understood here, is the ability to represent mental states such as beliefs, intents, and desires and attribute these to other individuals (Premack and Woodruff, 1978). This ability has been studied extensively by primatologists (Call and Tomasello, 2008), developmental psychologists (Wellman et al., 2001), autism researchers (Baron-Cohen et al., 2000) and, for the last decade or so, cognitive neuroscientists (cf. Saxe 2006 for a review and Spreng et al. 2008 for a recent meta-analysis). In this chapter we will report an analysis of TOM related brain activation during passive listening to narratives. While a full review of the relevant literature is well beyond the scope of this chapter (cf. Saxe et al. 2004 for a cross-field review and synthesis), we will address a number of critical issues before turning to our results. In particular, we will distinguish TOM from TOM dependent processing (4.1), discuss the distinction between the attribution of intentions and beliefs (4.2) and the delicate relationship between TOM and language (4.3). Finally, we will briefly address the growing field of TOM related neuro-imaging (4.4).
4.1 The competence performance distinction in the study of TOM

From an ecological (or evolutionary) perspective (Povinelli and Preuss, 1995), TOM can be viewed as an important resource for an individual needing to interact with other individuals. In effect, certain researchers define TOM as “the ability to explain and predict the behavior of conspecifics, based on observation of their intentional actions” (Ciaramidaro et al. 2007, p. 2007)\(^1\). The view of TOM as acts of inference and/or prediction probably stems from the experimental paradigms most often used to investigate this ability. One of the most prevalent paradigms is the ‘false belief task’ (Wimmer and Perner, 1983).

At the first stage of such experiment both the subject and a third person protagonist observe a certain action or fact (e.g., a cookie is being placed in a box). At the second stage the protagonist is absent and the subject alone witnesses a change to the original state of affairs (the cookie is moved to a bag). At the third and critical stage the subject is asked to predict the action of the protagonist related to the manipulated state of affairs (where would the protagonist look for the cookie?). The subject succeeds on the task if she predicts a course of action that reflects the (false) belief of the protagonist (in the box), rather than her (true) belief (in the bag). In this paradigm, the ability to represent another’s belief state is diagnosed by the ability of a subject to:

- Infer a state of knowledge or belief of a third person (belief not shared by the subject, and hence, false), usually through visual inspection of a scene or an event or based on a linguistic description of such an event
- Predict an action of that person based on that false belief

In order for a subject to succeed on such a task, the subject must be able to represent

---

\(^1\)Frith and Frith 2005 propose a similar definition: “We naturally explain people’s behavior on the basis of their minds: their knowledge, their beliefs and their desires . . . Explaining behavior in this way is called ‘having a theory of mind’ or ‘having an intentional stance’. ” (p.R644)
a belief other than her own, and hence the evidence for TOM. However, as pointed out by Bloom and German (2000), this task requires more than just TOM (as defined here). Among other things, it requires an inhibition mechanism to inhibit the ‘true belief’ response (Leslie et al., 2004) and more generally a decision procedure as well as an inference procedure or an algorithm to translate perceived action into action representations in terms of goals and beliefs and so on. Importantly, although all these additional components are critical for the success on a false belief task, they are not part of TOM per-se. We would like to suggest that the view of TOM cited above errs in failing to distinguish between a competence level description (the ability to represent other people’s mental states) and a performance level description (performance on tasks requiring such competence).

Complex tasks, such as the false belief task discussed above, are necessary in experimental paradigms that require a behavioral dependent measure. For example, success on a ‘true belief’ task could not be taken as evidence for TOM since the overt response based on the represented belief of a protagonist could not be distinguished from a response based on the subject’s own mental state. Brain imaging is not limited in that way since brain activation pattern could potentially disambiguate between behaviorally ambiguous conditions. In fact, as has been highlighted by a number of authors (Saxe et al., 2004), the use of complex tasks, such as the false belief task, in the context of brain imaging is complicated if one is interested in mapping cognitive processes onto brain anatomy. Since, as discussed above, such tasks contain a number of sub-tasks, interpretation of the associated brain activation requires many-to-many mapping, a non-trivial task.

Saxe and Powell (2006) (S&P) developed a novel fMRI paradigm for the study of TOM which makes use of the advantages of neuro-imaging while attempting to reduce the above mentioned confounds. Subjects passively read short descriptions of individuals. Only in one of the conditions did the descriptions include reference to the mental state of the individuals. In the context of the discussion here, it is of relevance to note that this paradigm does not require either inference (the mental state is explicitly reported) nor inhibition since no response was required and since
there was no false/true belief contrasts to begin with.

Though we will return to the results of this experiment in more details in the discussion section, it is worth while noting that as expected, this manipulation engaged only a subpart of the brain network observed to be active in the full fledged false belief task (which was also administered by the authors). In particular, the medial prefrontal cortex and anterior cingulate, which have been consistently implicated by brain imaging of false belief tasks (Gallagher and Frith, 2003; Spreng et al., 2008), did not respond differentially to TOM narratives compared to non-TOM ones suggesting that they do not play a direct role in TOM but rather support other dimensions of the complex false belief task.

Since in this chapter we will make use of (and expend on) S&P's paradigm, it is important to bear in mind the following point. Unlike classical TOM tasks, there is no behavioral evidence in this task that subjects are indeed engaging their TOM ability. The success of this paradigm is predicated on the assumption that simply processing linguistic information that pertains to other people's state of mind is sufficient to trigger TOM related brain activation. This assumption is, quite clearly, not self evident. However, we take the results from S&P's experiment to confirm it. Our experiment, and the results reported here, will presuppose the correctness of this assumption and will go further in assuming that manipulation of the type of linguistic information (intensions vs. belief states) triggers the corresponding type of mental representation.

4.2 Intentions vs. beliefs

Saxe et al. (2004), Tomasello et al. (2005), and Call and Tomasello (2008), reviewing the literature on primate and infant TOM research, all converge on an interesting conclusion. Primates, infants and children younger than 3 or 4 years old, are able to infer, represent, and act upon another's (often unobserved) goals or intentions. For example, in an imitation task, both chimpanzees and young children reproduce the (inferred) goal of the action but not necessarily the same movement, also when the
goal of the observed action was never achieved (due to perceived failure). Similarly, both infants and non-human primates distinguish accidental and intentional actions as well as inability vs. unwillingness to cooperate. For example, Woodward (1998), using an habituation paradigm, demonstrated that 6 month old infants encode a reaching event in terms of the desired goal rather than the specific path taken by the actor (after habituation to a particular reaching event, change of goal, but not change of path led to longer looking times at test, or dishabituation). Critically, this pattern was not observed when the human arm was replaced by an inanimate stick.

Buttelmann et al. (2007) showed chimpanzees how to operate a novel apparatus. The experimenters never used their hands (the natural body part to perform the action) in the demonstration, but rather some other body part (head, legs). In one condition the experimenter had her hands full (thus providing an 'explanation' for the use of the alternative body part). In the second condition the experimenter's hands were free. Just like young children (Gergely et al., 2002), when operating the apparatus themselves, the chimpanzee used their hands in the first condition but imitated the experimenters' body part choice in the second. In order to explain the different choice of body part in the two conditions, the experimenters argue that one must assume that the chimpanzee interpret the experimenter's behavior in terms of goal directed activity and not simply imitate the observed action. Imitation is ruled out by the fact that in the full hands condition the chimpanzee did not strictly imitate the experimenter but instead used its hands. This use of hands demonstrates that the chimpanzee's encoding of the event was not in term of specific actions but in terms of the goal (operate the apparatus). In view of the first condition, the fact that in the 'free hand' condition the chimpanzee did seem to imitate the choice of body part suggests that the chimpanzee reasoned that the experimenter, being a rational agent, would have used her hands (the most 'natural' choice) if she could. Since she chose to use a different body part, the chimpanzee concluded that hands were inappropriate for the operation of the novel apparatus (the 'goal') and so it followed the experimenter's choice of body part.

This ability to attribute intentionality to others and interpret other's behavior
in term of goal oriented action strongly argues that young children and non-human primates do possess some form of TOM as defined here. In contrast, they fail quite consistently on tasks such as the false belief task discussed above (though cf. Onishi and Baillargeon 2005).

Hare et al. (2001) tested chimpanzees on different variants of the false belief task. A subordinate chimpanzee had to approach a food item in the presence of a dominant chimpanzee. The dominant chimpanzee was either informed regarding the location of the food, uninformed, or misinformed (it saw the food in its original location but didn’t see it being moved). The subordinate chimpanzee’s behavior indicated that it could take advantage of the lack of knowledge by the other chimpanzee (the uninformed condition) but it did not take a similar advantage in the misinformed condition. Call and Tomasello, reviewing these results, interpret them to indicate that chimpanzees are able to represent ignorance of another regarding knowledge they do possess (this conclusion can also be made based on patterns of competitive food gathering in nature) but are unable to represent a false belief.

Apart from failure on the false belief task before the age of 3 or 4, Saxe and colleagues report that children start spontaneously producing genuine references to thought and belief only around 3 years old while reference to desire and emotions appears significantly earlier and it is not until 6 month later that they start explaining action using belief attribution. Wellman et al. (2001) reported the result of a longitudinal study that found a correlation between an infants’ (10-12 month) attention to a person’s intentions and the performance of the same child 3 years later on a battery of TOM and false belief tasks. This correlation suggests that despite the differences in development, the two aspects of TOM are related.

Having observed the dichotomy in the developmental literature between attribution of intentions and desires on the one hand and belief attribution on the other, Saxe and colleagues suggest that neuro-imaging could be used to address the relation between these two aspects of TOM:

\[\text{\footnotesize\textsuperscript{2}In normal circumstances of shared knowledge, a subordinate chimpanzee will not approach first food in the presence of a dominant chimpanzee}\]
Reasoning about beliefs develops later than ... an earlier theory of mind that includes attribution of desires, perceptions, and emotion. Does the later emerging competence colonize the same neural systems that underpin earlier reasoning? If so we would predict that attributions of desires, for instance, would recruit the same brain regions identified above as involved in belief attribution. If, on the other hand, reasoning about beliefs draws on distinct systems or abilities, then desire attribution should not produce activity in regions associated with belief attribution, and may recruit a distinct set of brain regions. (p. 108)

The authors report results from a number of imaging studies regarding the processing and representation of goal directed action or intentions but point out the absence of direct comparisons with belief attribution. In addition, the study of action understanding also often considers the competence question (the representation of intention) to be identical to the performance question (how we infer intentions from observation of actions). For example, a recent attempt to address the question posed by Saxe and colleagues (Gobbini et al., 2007) compared fMRI correlates of a task of inferring social intentions from animated clips to correlates of a false belief task. The authors found the two tasks to induce largely non-overlapping activations. However, since both tasks differ on multiple sub-components apart from the crucial TOM distinction, these results are extremely hard to interpret. Saxe and Powell (2006) used a passive reading paradigm (discussed above) to compare descriptions containing belief attribution descriptions of non-mentalistic internal bodily sensations (hunger, fever). We were not able to find in the (very vast and expanding) literature on the topic, an

---

3In fact, in the domain of action observation the discovery of mirror neurons in primates (neurons that fire both for an execution of an action and the observation of the same action performed by another individual; see Iacoboni and Mazziotta 2007 for a recent review) has been used to support an explicit reductive approach for intention. According to this view, intention and action are represented by one and the same system (Fogassi et al., 2005). However, other researchers have demonstrated that at least explicit attention to others' intentions activates brain regions not associated with the representation of action per-se (de Lange et al., 2008). A more general problem with this line of research is that intention and desire are not restricted to the domain of executable action.
attempt to use a similar design to compare the attribution of intentions or desires to that of belief.

4.3 TOM and language

Multiple sources of evidence, in particular from (a)typical development, point to an intimate relation between language (or at least language development) and TOM (or more specifically, success on the false belief task) (Milligan et al., 2007). Numerous linguistic performance measures correlate with success on the TOM task, such as receptive vocabulary, expressive narrative speech and receptive measures of syntax and semantics, even after age is factored out. Furthermore, longitudinal studies have shown that language abilities at a younger age predict TOM performance at a later age (cf. Slade and Ruffman 2005 for relevant references and discussion).

While this relation is by now well established, researchers disagree what constitutes the explanation for this relation. TOM, and in particular the ability to interpret intention appears to be critical for early word acquisition (Baldwin and Moses, 2001) so maybe the development of TOM supports language acquisition (Bloom, 2000). On the other extreme, maybe language is the representational medium for TOM (and in particular belief attribution) and so linguistic development is crucial for children to reason about other people’s beliefs (de Villiers and de Villiers, 2000). A third approach suggests that language is a pre-requisite for the activity of conversation, which is the medium through which children are exposed to other people’s beliefs (Nelson, 2005). Finally, and along the lines of our discussion earlier, language might be relevant to the specific task (false belief) but for reasons other than TOM (Bloom and German, 2000). For example, Miller (2001), who tested children with Specific Language Impairment (SLI), found them to perform as well as age-match control on a version of the false belief task that used simpler syntactic structure (but still required false belief attribution). In a version of the task which employed more articulated syntax, the same children under-performed in comparison to controls.

Evidence from late signers (deaf children that were not exposed to sign language
before the age of 6 or 7, de Villiers 2005; Meristo et al. 2007) has provided strong support for the view that takes language development to drive TOM development (or success on the false belief task). Late signers are significantly delayed on both signed and non-verbal versions of the false belief task. However, these findings are compatible both with the view that language is a necessary for the representation of TOM and the view according to which language is necessary in order to help children learn about other people’s minds.

Behavioral data on the relation between theory of mind and language in adults is more sparse. A case study by Varley and Siegal 2000 of an agrammatic aphasic that was able to pass the false belief task is often cited as an evidence that TOM is independent of language in the adult state. While this finding is of great interest it is important to keep in mind that agrammatism, even of the severe kind described by these authors (with both syntactic production and comprehension severely impaired), indicates that a necessary brain tissue for language production has been compromised but not necessarily that all of the language architecture has been destroyed. It is at least conceptually possible that the part of the linguistic system that underlies TOM reasoning is still intact (though due to other damage to the system, does not surface in production).

Newton and de Villiers (2007) had college students perform a non-verbal false belief task while verbally shadowing (repeating speech presented over headphones). This paradigm has been used in the past to investigate the role of language in apparently non linguistic tasks such as spatial navigation (Hermer-Vazquez et al., 1999). Verbal shadowing caused college students to fail on the false belief task. Performance on the same task was not affected by an equally hard rhythm shadowing task. These results indicate a role for language in reasoning about false beliefs (and possibly TOM more generally) even in the adult state and so argue against the hypothesis that language serves only to facilitate the acquisition of TOM.

An orthogonal question that remains to be answered is what specific aspect of language, if any, is linked to TOM. J. De Villiers and colleagues (de Villiers and de Villiers, 2000; de Villiers and Pyers, 2002) have advanced the view that the aspect
of the linguistic system most important for TOM is the syntactic representation of tensed\textsuperscript{4} sentential complements (as in *John believes that Mary bought a new book*).

Slade and Ruffinan (2005), testing 3.8 years old kids twice (with a 6 month interval), did not find any specific aspect of language to be more predictive of success in a false belief task than any other (though overall language ability did predict TOM results). Milligan et al. (2007) performed a meta-analysis of studies that examined the relation between language and TOM development. Out of over 100 studies, only 4 tested a specific hypothesis (the sentential complements proposal mentioned above) about the linguistic mechanism that enables TOM. Most other studies used clinical general language assessment tests. Certain studies did use tests of a more specific language ability but without explicating a linking hypothesis. The meta-analysis found that all language measures (divided into receptive vocabulary, general language, syntax, semantics and sentential complements) had a significant contribution to the explanation of TOM performance. Though numerically the effect of sentential complements was the strongest, due to the small number of studies it did not come out significantly different that other measures. The only measure that had significantly less explanatory weight than all other measures was the receptive vocabulary measure.

It is very hard to draw any conclusions from the current state of the art. On the one hand, the only specific hypothesis on the market (De Villiers’ tensed complements) is itself quite complex and moreover might involve some circularity (cf. Slade and Ruffinan 2005). On the other hand, observations that general language level (or broad terms such as syntax or semantics) correlate with TOM cannot be said to argue against a more specific relation (since it could be that the more specific relation drives the more general effect) without specific testing\textsuperscript{5}. As a consequence, apart from the conclusion that (some aspect of) language has an effect on TOM performance across

\textsuperscript{4}De Villiers restricts the relevant linguistic structure to tensed complements in English presumably because in English realis mood requires tense.

\textsuperscript{5}Slade and Ruffinan (2005) represents an effort in this direction, since the authors tested relatively more specific aspects of language ability alongside general measures. The problem is that they did not offer a linking hypothesis between these specific aspects of language (e.g. word order) and TOM.
studies, there is little more we can conclude at the moment. One would hope that
future experiments will test more specific hypotheses regarding the relation between
language and TOM.

4.4 A note on neuro-imaging of TOM

Echoing the perspective of Saxe et al. (2004), neuro-imaging offers a new and exciting
window on the nature of internal representations such as TOM. A recent meta-analysis
(Spreng et al., 2008) found over 130 imaging (PET and FMRI) studies between 1985
and June 2007. Many of these papers, in particular those of earlier years, replicated
the behavioral false belief paradigm. Other papers (cf. Saxe and Powell 2006 dis-
cussed above) have taken a greater advantage of the methodology and have abandoned
the false belief task in favor of more ‘passive’ (or naturalistic) tasks such as narrative
comprehension or simulated games.

More importantly, neuro-imaging has permitted the field to investigate the rep-
resentation of TOM and its relation to other cognitive domains without the need to
rely on a correlated deficit logic (as was the case in the behavioral paradigm). A
consistent network of brain regions has emerged from these multiple experiments (cf.
Spreng et al. (2008) for a meta-analysis) investigating TOM. This network includes
the left and right temporal parietal junction (TPJ), regions within the precuneus
and posterior cingulate, the bilateral temporal poles and adjacent anterior temporal
regions, the medial prefrontal cortex and anterior cingulate and the amygdala. At
the same time, not all papers identify all these regions. Moreover, other regions are
often implicated as well (Spreng and colleagues’ meta-analysis identified 20 clusters,
at least 8 of which are not part of the network mentioned above).

In view of this very distributed network and the flexibility offered by brain imaging,
many new (and old) hypotheses regarding the representation of TOM have been put
to the test in an attempt to explain the role of the different components of the
networks. One kind of experiment involves, first, the identification of another domain
that might share mechanisms with, or underly, TOM processing. Then, two tasks are
designed, one which has a clear TOM character (belief or intention attribution) and another which does not have an explicit TOM component but does implicate the second domain. Since most of these studies use a subtractive logic, baseline tasks are also included. Sometime the same baseline task is used for both target conditions but sometimes different baselines are used.

Brain activation (target condition-baseline) are compared across conditions and sometimes direct subtraction across the two conditions is performed as well. Cognitive domains that have been explored in relation to TOM include logical reasoning (Ferstl and von Cramon, 2002; Fiddick et al., 2005; Ermer et al., 2006), biological movement (Saxe et al., 2004), spatial reorientation (Corbetta et al., 2008), episodic memory (Rosenbaum et al., 2007), action understanding (de Lange et al., 2008) and emotion (Hynes et al., 2006).

A second kind of experiment proposes a distinction between different types, mechanisms or aspects of TOM proper (or TOM related tasks) that might be supported by different brain structures within the network (or even split the large network into smaller, complementary ones). H. Walter and colleagues Walter et al. (2004); Ciaramidaro et al. (2007) have proposed to distinguish between social or common intentions, private intentions and communicative intentions. Saxe and Powell (2006) contrasted representation of a third party’s mental attitudes to the representation of other internal states. Hamilton and Grafton (2006) examined goal representations. Mitchell et al. (2005) manipulated the extent of self projection. The role of empathy in TOM was studied by Hynes et al. (2006); Vllm et al. (2006) and den Ouden et al. (2005) tease apart prospective memory from causal inferences. While this emerging field of study presents much potential, an informal survey of the existing literature does not yet suggest a convergence on a finer grained description of the TOM network.

4.5 The current experiment

We had three major goals in mind with respect to the question of TOM representation in the context of the experiment reported here.
4.5.1 Identification of the TOM network

First, we wanted to find out whether it is possible to identify the TOM network in a naturalistic passive listening task that did not require explicit TOM manipulation. In addition, our narratives were not designed to contain particularly taxing or complex TOM scenarios and in that respect are different from false belief scenarios. If the TOM network would be activated under these conditions, the result will strengthen the conclusion that it is truly implicated in TOM per-se rather than an artifact of the task complexity or other particular feature of false belief scenarios.

Earlier work using narrative comprehension to study TOM (cf. Saxe and Powell 2006; Saxe and Wexler 2005; Ferstl and von Cramon 2002, among others) used a traditional block design which allowed averaging of the signal across blocks of the same condition thus increasing signal to noise ratio. A central feature of the paradigm used here is the choice to abandon this traditional design in favor of a direct modeling of the entire narrative event as a single, complex, task. In our design there is no clear temporal separation between different conditions. In that respect, given the well established anatomical characterization of the TOM network, TOM can be used as a test-case for this novel paradigm. Since this first goal was essentially to replicate existing results we used a variation on the coding scheme in Saxe and Powell (2006). Each sentence was used a variation on the coding scheme in Saxe and Powell (2006).

- 3: Sentences that pertain to the mental state of the protagonists.
- 2: Sentences that pertain to humans but not to their mental state.
- 1: Sentences that do not pertain to humans or where all human reference is backgrounded.

These values were entered (mean corrected) as parametric modulators in the sentence design, (where each sentence was modeled as a block) which is conceptually similar.

---

6 We thank Rebecca Saxe for suggesting us to code our narratives for TOM and for her continuous support and useful advice.

7 The coding in this case was done by the author following guidelines provided by Rebecca Saxe.
to the paradigm of Saxe and Powell (2006). We expected activation in the TOM network to show positive correlation with this regressor replicating the results of Saxe and Powell as well as others.

4.5.2 Distinguishing intentions and beliefs

In an attempt to address the question regarding the relation between intention/desire and belief attribution (cf. Saxe et al. 2004 and 4.2) we added 2 new parametric modulators to our word model. In the intentionality parameter, we assigned 1 to all the words in the narratives that made reference to the desire, intention or wish of a protagonist (mostly verbs such as want or hope, wait for but also nouns such as goal or purpose) and 0 to all other words. In the belief parameter, we assigned 1 to all words that made reference to a protagonist’s beliefs, knowledge or thought processes (verbs such as think, believe, know and nouns such as fact, idea, belief).

The assumption behind these regressors is an extension of the logic in Saxe and Powell and in our sentence level design. Namely, we assume that the processing of intentional content should activate brain regions that engage in the representation of others’ intentions and similarly for belief states. This design makes another important assumption in taking the occurrence of certain lexical items as a proxy for TOM content (this is different from the sentence model where TOM value was a function of sentence meaning).

The use of lexical items as proxies raises the possibility of a confound brought to our attention by D. Pesetsky and D. Fox (p.c.). Since each of our word level TOM regressors track the presence of certain class of words, how can we distinguish between activation related to TOM and activation due to some common lexical characteristic of the word class? There is no straightforward answer to this concern. It is worthwhile to note that the words included in each regressor were not restricted to a specific grammatical class (combining verbs, nouns and adjectives). In addition, if the shared feature of all these words is related to TOM, then this is not a confound but a relevant finding (since we are not attempting, by this contrast, to distinguish linguistic and non linguistic aspects of TOM). Finally, as we will see in the results section, one
can sometimes use prior knowledge regarding the implicated brain regions to infer whether the effect is due to the lexicon.

We were interested to find out to what extent the intentionality measure activated the same network as the belief state measure or whether other regions (such as the action representation network) or other sub-parts of the TOM network were implicated for the representation of intentions. We produced a main effect contrast map for each of these parameters as well as a difference map (belief-intentionality).

4.5.3 The relation between modality and TOM

From a linguistic or philosophy of language perspective, most of the terms used to make reference to TOM belong to the larger class of modal expressions. von Fintel defines modality in the following manner:

Modality is a category of linguistic meaning having to do with the expression of possibility and necessity. A modalized sentence locates an underlying or prejacent proposition in the space of possibilities.... *Sandy might be home* says that there is a possibility that Sandy is home. *Sandy must be home* says that in all possibilities, Sandy is home. The counterpart of modality in the temporal domain should be called temporality, but it is more common to talk of tense and aspect, the prototypical verbal expressions of temporality. Together, modality and temporality are at the heart of the property of displacement (one of Charles F. Hocketts design features of human language) that enables natural language to talk about affairs beyond the actual here and now. (von Fintel, 2006)

Different types of modality express different paths of "displacement" (or different accessibility relations in possible worlds talk). For example, *must*, as used in von Fintel's example, is an epistemic modal since it describes what is necessarily the case given the speakers knowledge of the world. *Must* can also be used to express other modalities, as in *Sandy must pay a fine*, where it is used deontically to describe what is necessarily the case given the law. *Alethic* or logical modality expresses general
truths which are independent of the actual world (*a triangle must have three sides*).

The TOM expressions we included in the parameters described above fall broadly into two categories. The intentional TOM terms (such as *want* or *desire*) are known as *bouletic* modals since they describe what is necessary or possible given the subject’s *desires*. Verbs such as *believe* are doxastic modals, describing what is possible or necessary given the subject’s belief (or knowledge) about the world. One question, usually not addressed by the current imaging literature, is what role, if any, modality plays in the activation of the TOM network.

Most often, TOM scenarios are compared to non-modal ones (rather than to non-TOM modal scenarios). It is possible that at least part of the TOM specific activation reported by these experiments should in fact be attributed to the more general ‘displacement capacity which underlies modality. In fact, a recent review of the default network (Buckner et al., 2008) has suggested a role for that network (which overlaps with the TOM network as well as a number of others) in terms not dissimilar to the ones employed here to describe modality:

Remembering, envisioning the future, and conceiving the mental states of others … all converge on similar core processes. In each instance, one is required to simulate an alternative perspective to the present. These abilities, which are most often studied as distinct, rely on a common set of processes by which mental simulations are used adaptively to imagine events beyond those that emerge from the immediate environment. (p. 23)

This convergence between linguistic and neuro-cognitive theory has motivated us to investigate the relation between TOM and more general modal cognition. In order to operationalize this comparison we added a modal parameter to the model, in which we assigned 1 to all the words in the narratives that contributed a modal meaning to the exclusion of items that fell into either the intentionality or belief categories. These parameters included an heterogeneous group of modal expressions such as ability modals (*can, able*), necessity (*must, have*), epistemic possibility (*might,
could), conditionals (if) and the future modal (will). We were interested in the main effect of the modal and in particular in the contrast TOM-modal (where TOM stands for the combination of intentionality and belief).

4.5.4 TOM and syntax

Even if natural language syntax is not the medium of representation of TOM as has been proposed by Collins (2000); de Villiers (2007), it is still the case that linguistically mediated TOM descriptions usually require more articulated syntax and in particular sentential complementation. Under the perspective that distinguishes TOM from language, this is a potential source of confound. For example, In the sample stimuli presented by Saxe and Powell (2006) only the TOM narratives contained sentential complementation (or S recursion). One way to avoid this confound is by equating syntactic complexity across conditions. In our experiment we use instead concurrent modeling of syntactic surprisal and number of derivation steps. This design choice allows us, in addition, to address the more general question regarding the relationship between TOM and syntax. Specifically we can compare the activation maps of these two parametric modulations to find out any overlaps. We will return to this comparison in the general discussion.

4.6 Results

4.6.1 Sentence level effects of TOM

Peaks of positive correlation with the sentential TOM measure ($P < 0.005$ uncorrected, cluster threshold =20) are reported in table (4.1).
Table 4.1: TOM Results

<table>
<thead>
<tr>
<th>Region</th>
<th>Location</th>
<th>MNI coordinates (x,y,z)</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>L temporal</td>
<td>posterior superior temporal sulcus</td>
<td>-46 -64 34</td>
<td>6.0523</td>
</tr>
<tr>
<td></td>
<td>middle temporal gyrus</td>
<td>-62 -12 -26</td>
<td>4.1799</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-66 -38 -6</td>
<td>4.0269</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-60 -20 -16</td>
<td>3.7903</td>
</tr>
<tr>
<td></td>
<td>fusiform</td>
<td>-30 -54 -8</td>
<td>3.8069</td>
</tr>
<tr>
<td>L frontal</td>
<td>anterior superior frontal gyrus</td>
<td>-2 54 32</td>
<td>5.9372</td>
</tr>
<tr>
<td></td>
<td>middle frontal gyrus</td>
<td>-40 16 50</td>
<td>5.0083</td>
</tr>
<tr>
<td></td>
<td>superior frontal gyrus</td>
<td>-4 32 62</td>
<td>4.6321</td>
</tr>
<tr>
<td></td>
<td>orbital part</td>
<td>-50 30 -10</td>
<td>4.3896</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-42 34 -10</td>
<td>3.7062</td>
</tr>
<tr>
<td></td>
<td>rectus gyrus</td>
<td>-2 46 -20</td>
<td>4.3009</td>
</tr>
<tr>
<td></td>
<td>orbital part</td>
<td>-34 16 -20</td>
<td>3.98</td>
</tr>
<tr>
<td>L parietal</td>
<td>angular gyrus</td>
<td>-58 -56 30</td>
<td>6.1854</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-48 -52 36</td>
<td>3.8534</td>
</tr>
<tr>
<td></td>
<td>precuneus</td>
<td>-6 -60 32</td>
<td>4.9305</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-50 30 -10</td>
<td>5.0292</td>
</tr>
<tr>
<td>L insula</td>
<td>posterior insula</td>
<td>-30 -22 2</td>
<td>4.7966</td>
</tr>
<tr>
<td>L subcortical</td>
<td>caudate</td>
<td>-12 12 8</td>
<td>4.948</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-8 0 2</td>
<td>3.797</td>
</tr>
<tr>
<td>R frontal</td>
<td>anterior superior frontal gyrus</td>
<td>-20 -20 2</td>
<td>4.1357</td>
</tr>
<tr>
<td></td>
<td>middle frontal gyrus</td>
<td>8 54 18</td>
<td>5.0907</td>
</tr>
<tr>
<td></td>
<td>orbital part</td>
<td>8 66 22</td>
<td>4.9912</td>
</tr>
<tr>
<td></td>
<td>superior frontal gyrus</td>
<td>42 16 38</td>
<td>4.0171</td>
</tr>
<tr>
<td></td>
<td></td>
<td>42 28 -2</td>
<td>3.4934</td>
</tr>
<tr>
<td>R parietal</td>
<td>superior frontal gyrus</td>
<td>42 -1 60</td>
<td>4.8208</td>
</tr>
<tr>
<td></td>
<td>precuneus</td>
<td>34 -1 68</td>
<td>4.7449</td>
</tr>
<tr>
<td></td>
<td>angular gyrus</td>
<td>16 30 62</td>
<td>4.3493</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 -56 30</td>
<td>4.8088</td>
</tr>
<tr>
<td></td>
<td>intraparietal sulcus</td>
<td>46 -56 32</td>
<td>3.9969</td>
</tr>
<tr>
<td></td>
<td>precuneus</td>
<td>56 -56 24</td>
<td>3.3751</td>
</tr>
<tr>
<td>R cingulate</td>
<td>dorsal cingulate</td>
<td>22 -54 50</td>
<td>3.9899</td>
</tr>
<tr>
<td>R sub-cortical</td>
<td>caudate</td>
<td>18 -58 38</td>
<td>3.125</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10 -28 26</td>
<td>3.9833</td>
</tr>
<tr>
<td></td>
<td></td>
<td>14 6 14</td>
<td>5.3377</td>
</tr>
<tr>
<td></td>
<td></td>
<td>14 14 8</td>
<td>4.142</td>
</tr>
</tbody>
</table>
The sentential TOM measure had a significant effect bilaterally in the angular gyrus (TPJ), inferior and dorsal frontal cortex (4-1.A,C), the anterior (and anterior ventral) prefrontal cortex and the precuneus/posterior cingulate (4-1.B,D). A left lateralized effect was observed in the mid temporal gyrus and the fusiform. Subcortical structures that exhibited significant effects included the bilateral caudate and left thalamus.
Figure 4-1: A: Lateral view of positive correlation with sentential TOM in the left hemisphere (red clusters in the TPJ, along the middle temporal gyrus, ventral prefrontal cortex, and superior prefrontal cortex). B: Medial view of positive correlation with sentential TOM in the left hemisphere (red clusters in the precuneus, fusiform gyrus, anterior prefrontal cortex, and ventral ventro-medial prefrontal cortex). C: Lateral view of positive correlation with sentential TOM in the right hemisphere (red cluster in the TPJ and lateral parietal lobe, dorsal and ventral frontal cortex). D: Medial view of positive correlation with sentential TOM in the right hemisphere (red clusters in the precuneus, dorsal cingulate and anterior prefrontal cortex). All data presented on a two dimensional representation of an 'inflated' cortex. The group t map was thresholded at $p < 0.005$, uncorrected, cluster threshold=20. Blue circles represent approximate location of peaks of the major clusters from the meta-analysis in Spreng et al. (2009).
4.6.2 Brain correlates of intentions and belief states

Peaks of positive correlation with the belief state TOM measure ($P < 0.005$ uncorrected, cluster threshold =20) are reported in table (4.2).

<table>
<thead>
<tr>
<th>Region</th>
<th>Location</th>
<th>MNI coordinates (x,y,z)</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>L temporal</td>
<td>anterior middle temporal gyrus</td>
<td>-46 8 -40</td>
<td>4.4072</td>
</tr>
<tr>
<td></td>
<td>middle temporal gyrus</td>
<td>-58 -6 -26</td>
<td>3.2403</td>
</tr>
<tr>
<td></td>
<td>inferior frontal sulcus</td>
<td>-66 -30 -12</td>
<td>5.9771</td>
</tr>
<tr>
<td>L frontal</td>
<td>anterior superior frontal gyrus</td>
<td>-20 58 34</td>
<td>5.1825</td>
</tr>
<tr>
<td></td>
<td>middle frontal gyrus</td>
<td>-30 28 50</td>
<td>4.2718</td>
</tr>
<tr>
<td></td>
<td>paracentral gyrus</td>
<td>-36 22 44</td>
<td>3.3847</td>
</tr>
<tr>
<td>L parietal</td>
<td>angular gyrus</td>
<td>-48 -64 36</td>
<td>6.321</td>
</tr>
<tr>
<td></td>
<td>inferior parietal</td>
<td>-58 -62 28</td>
<td>5.6117</td>
</tr>
<tr>
<td></td>
<td>precuneus</td>
<td>-6 -48 30</td>
<td>4.2892</td>
</tr>
<tr>
<td>L cingulate</td>
<td>posterior cingulate</td>
<td>-20 -32 36</td>
<td>4.4264</td>
</tr>
<tr>
<td></td>
<td>dorsal cingulate</td>
<td>-4 -20 36</td>
<td>3.8351</td>
</tr>
<tr>
<td>R temporal</td>
<td>temporal pole</td>
<td>48 4 -42</td>
<td>6.0771</td>
</tr>
<tr>
<td></td>
<td>middle temporal gyrus</td>
<td>66 -28 -14</td>
<td>5.0477</td>
</tr>
<tr>
<td></td>
<td>fusiform</td>
<td>34 -62 -10</td>
<td>4.1727</td>
</tr>
<tr>
<td></td>
<td>inferior temporal gyrus</td>
<td>42 -20 -14</td>
<td>4.0786</td>
</tr>
<tr>
<td></td>
<td>middle frontal gyrus</td>
<td>58 -18 -26</td>
<td>3.5819</td>
</tr>
<tr>
<td>R frontal</td>
<td>superior frontal gyrus</td>
<td>14 40 56</td>
<td>5.2455</td>
</tr>
<tr>
<td></td>
<td>anterior superior frontal gyrus</td>
<td>16 26 60</td>
<td>5.0501</td>
</tr>
<tr>
<td></td>
<td>superior frontal gyrus</td>
<td>5 44 30</td>
<td>5.0341</td>
</tr>
<tr>
<td></td>
<td>middle frontal gyrus</td>
<td>4 54 18</td>
<td>4.1435</td>
</tr>
<tr>
<td></td>
<td>inferior frontal gyrus</td>
<td>0 42 50</td>
<td>4.8563</td>
</tr>
<tr>
<td></td>
<td>superior frontal gyrus</td>
<td>46 16 54</td>
<td>3.8839</td>
</tr>
<tr>
<td></td>
<td>suprar-marginal gyrus</td>
<td>42 20 42</td>
<td>3.7578</td>
</tr>
<tr>
<td>R parietal</td>
<td>angular gyrus</td>
<td>58 -56 32</td>
<td>5.6708</td>
</tr>
<tr>
<td></td>
<td>suprar-marginal gyrus</td>
<td>44 -36 28</td>
<td>3.7195</td>
</tr>
<tr>
<td></td>
<td>angular gyrus</td>
<td>44 -46 30</td>
<td>2.9792</td>
</tr>
<tr>
<td>R occipital</td>
<td>precuneus</td>
<td>6 -52 26</td>
<td>4.6468</td>
</tr>
<tr>
<td></td>
<td>occipital pole</td>
<td>26 -88 -4</td>
<td>3.3673</td>
</tr>
<tr>
<td></td>
<td></td>
<td>28 -98 -6</td>
<td>3.0489</td>
</tr>
<tr>
<td></td>
<td></td>
<td>28 -92 4</td>
<td>2.9405</td>
</tr>
</tbody>
</table>
The belief state measure had a significant bilateral effect in the angular gyrus (TPJ), mid and anterior temporal gyrus, dorsal frontal cortex (4-2.A,C), precuneus and anterior prefrontal cortex (4-2.B,D). Left lateralized activation was observed in the posterior/dorsal cingulate. Right lateralized activation was observed in the temporal pole and the occipital pole.

Peaks of positive correlation with the intentional TOM measure ($P < 0.005$ uncorrected, cluster threshold =20) are reported in table (4.3).

<table>
<thead>
<tr>
<th>Region</th>
<th>Location</th>
<th>MNI coordinates (x,y,z)</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>L parietal</td>
<td>angular gyrus</td>
<td>-46 -62 30</td>
<td>5.6863</td>
</tr>
<tr>
<td>L frontal</td>
<td>anterior superior frontal gyrus</td>
<td>-8 60 36</td>
<td>5.6116</td>
</tr>
<tr>
<td></td>
<td>superior frontal gyrus</td>
<td>-10 32 54</td>
<td>3.5484</td>
</tr>
<tr>
<td>R frontal</td>
<td>superior frontal gyrus</td>
<td>16 38 48</td>
<td>4.9446</td>
</tr>
<tr>
<td>R temporal</td>
<td>anterior inferior temporal gyrus</td>
<td>46 -2 -36</td>
<td>4.5813</td>
</tr>
<tr>
<td></td>
<td>temporal pole</td>
<td>48 14 -34</td>
<td>4.3435</td>
</tr>
<tr>
<td>R frontal</td>
<td>anterior superior frontal gyrus</td>
<td>12 64 12</td>
<td>3.6934</td>
</tr>
<tr>
<td>R parietal</td>
<td>angular gyrus</td>
<td>58 -60 28</td>
<td>3.5817</td>
</tr>
</tbody>
</table>

The intentional measure had a significant bilateral effect in the angular gyrus (TPJ) and dorsal prefrontal cortex (4-2.A,C). Right lateralized effect was observed in the temporal pole.

Peaks of positive difference between the TOM measures (belief states > intentions) ($P < 0.005$ uncorrected, cluster threshold =20) are reported in table (4.4). Peaks of positive difference between the intention and TOM measures (intentions > belief states) ($P < 0.005$ uncorrected, cluster threshold =20) are reported in table (4.5).
### Table 4.4: TOM Results: beliefs - intentions

<table>
<thead>
<tr>
<th>Region</th>
<th>Location</th>
<th>MNI coordinates (x,y,z)</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>L temporal</td>
<td>superior temporal</td>
<td>-42 -28 6</td>
<td>3.9203</td>
</tr>
<tr>
<td>L frontal</td>
<td>anterior superior frontal gyrus</td>
<td>-8 52 18</td>
<td>6.6131</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-18 52 14</td>
<td>5.1016</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-24 58 12</td>
<td>3.6003</td>
</tr>
<tr>
<td></td>
<td>orbital</td>
<td>-46 18 -22</td>
<td>3.5416</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-30 22 -16</td>
<td>3.5089</td>
</tr>
<tr>
<td>L parietal</td>
<td>anterior superior frontal gyrus</td>
<td>-4 52 -4</td>
<td>3.3347</td>
</tr>
<tr>
<td></td>
<td>angular gyrus</td>
<td>-48 -50 38</td>
<td>5.1012</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-46 -42 30</td>
<td>3.8072</td>
</tr>
<tr>
<td></td>
<td>postcentral sulcus</td>
<td>-38 -24 30</td>
<td>4.0198</td>
</tr>
<tr>
<td></td>
<td>angular gyrus</td>
<td>-44 -30 30</td>
<td>4.576</td>
</tr>
<tr>
<td>L cingulate</td>
<td>dorsal cingulate</td>
<td>-2 -24 32</td>
<td>3.4414</td>
</tr>
<tr>
<td></td>
<td>anterior cingulate</td>
<td>-4 -16 34</td>
<td>3.4065</td>
</tr>
<tr>
<td>L insula</td>
<td>anterior insula</td>
<td>-38 18 -14</td>
<td>4.0391</td>
</tr>
<tr>
<td>R temporal</td>
<td>inferior temporal gyrus</td>
<td>54 -24 -28</td>
<td>4.0131</td>
</tr>
<tr>
<td></td>
<td>middle temporal gyrus</td>
<td>62 -26 -12</td>
<td>3.7434</td>
</tr>
<tr>
<td></td>
<td>superior temporal sulcus</td>
<td>58 -32 -8</td>
<td>3.7278</td>
</tr>
<tr>
<td>R frontal</td>
<td>superior frontal gyrus</td>
<td>8 30 54</td>
<td>4.4766</td>
</tr>
<tr>
<td>R parietal</td>
<td>angular gyrus</td>
<td>50 -42 40</td>
<td>3.5847</td>
</tr>
<tr>
<td>R cingulate</td>
<td>anterior cingulate sulcus</td>
<td>40 -38 36</td>
<td>3.4145</td>
</tr>
<tr>
<td>R</td>
<td>hippocampus</td>
<td>12 42 16</td>
<td>4.2322</td>
</tr>
<tr>
<td></td>
<td>hippocampus/amygdala</td>
<td>38 -26 -12</td>
<td>4.8371</td>
</tr>
</tbody>
</table>

### Table 4.5: TOM Results: intentions - beliefs

<table>
<thead>
<tr>
<th>Region</th>
<th>Location</th>
<th>MNI coordinates (x,y,z)</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>L sub-cortical</td>
<td>parahippocampus</td>
<td>-28 -24 -26</td>
<td>4.2214</td>
</tr>
<tr>
<td></td>
<td>hippocampus</td>
<td>-18 -8 -20</td>
<td>3.9899</td>
</tr>
<tr>
<td></td>
<td>hippocampus/amygdala</td>
<td>-22 -2 -26</td>
<td>3.3547</td>
</tr>
</tbody>
</table>
The contrast between belief states and intentions had a bilateral effect in the anterior portion of the angular gyrus (4-2.A,C) and the medial anterior prefrontal cortex (4-2.B,D). A left lateralized effect was observed in the inferior frontal gyrus (orbital part) and the posterior cingulate. A right lateralized effect was observed in the mid temporal gyrus. The opposite contrast (intentions-beliefs) had a significant effect in the left hippocampus and parahippocampus.
Figure 4-2: A: Lateral view of positive correlation with the two word level TOM measures in the left hemisphere. B: Medial view of positive correlation with the two word level TOM measures in the left hemisphere. C: Lateral view of positive correlation with the two word level TOM measures in the right hemisphere. D: Medial view of positive correlation with the two word level TOM measures in the right hemisphere. All data presented on a two dimensional representation of an 'inflated' cortex. The group t map was thresholded at $p < 0.005$, uncorrected, cluster threshold=20. Belief states in red, intentions in yellow, overlap in green. The contrast belief-intention is displayed in purple. The sentential TOM measure is reproduced in light blue for reference.
### 4.6.3 Brain correlates of modality and TOM

Peaks of positive correlation with the modal measure ($P < 0.005$ uncorrected, cluster threshold $= 20$) are reported in table (4.6).

<table>
<thead>
<tr>
<th>Region</th>
<th>Location</th>
<th>MNI coordinates (x,y,z)</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>L frontal</td>
<td>superior frontal gyrus</td>
<td>-10 -6 70</td>
<td>4.3187</td>
</tr>
<tr>
<td></td>
<td>central sulcus</td>
<td>-60 -4 28</td>
<td>3.9538</td>
</tr>
<tr>
<td></td>
<td>superior frontal gyrus</td>
<td>-54 -12 28</td>
<td>3.3141</td>
</tr>
<tr>
<td></td>
<td>inferior parietal</td>
<td>-18 -14 64</td>
<td>3.8524</td>
</tr>
<tr>
<td>L parietal</td>
<td>anterior insula</td>
<td>-18 -14 64</td>
<td>3.8524</td>
</tr>
<tr>
<td>L insula</td>
<td>superior frontal gyrus</td>
<td>-36 10 -8</td>
<td>4.7525</td>
</tr>
<tr>
<td></td>
<td>anterior insula</td>
<td>-34 0 10</td>
<td>4.1913</td>
</tr>
<tr>
<td></td>
<td>central sulcus</td>
<td>-54 -36 56</td>
<td>4.0561</td>
</tr>
<tr>
<td></td>
<td>inferior parietal</td>
<td>-60 -36 56</td>
<td>3.6311</td>
</tr>
<tr>
<td>L sub-cortical</td>
<td>putamen</td>
<td>-16 6 -8</td>
<td>4.4342</td>
</tr>
<tr>
<td>R frontal</td>
<td>middle frontal sulcus</td>
<td>20 -6 46</td>
<td>6.0134</td>
</tr>
<tr>
<td></td>
<td>opercular part</td>
<td>56 10 2</td>
<td>3.9952</td>
</tr>
<tr>
<td></td>
<td>superior frontal gyrus</td>
<td>48 8 0</td>
<td>3.7483</td>
</tr>
<tr>
<td></td>
<td>central sulcus</td>
<td>66 -6 22</td>
<td>4.0889</td>
</tr>
<tr>
<td></td>
<td>precentral sulcus</td>
<td>60 -2 26</td>
<td>3.0261</td>
</tr>
<tr>
<td>R parietal</td>
<td>supra-marginal gyrus</td>
<td>56 4 18</td>
<td>3.745</td>
</tr>
<tr>
<td></td>
<td>intraparietal sulcus</td>
<td>64 -30 46</td>
<td>5.7406</td>
</tr>
<tr>
<td></td>
<td>cingulate sulcus</td>
<td>48 -34 32</td>
<td>4.3233</td>
</tr>
<tr>
<td></td>
<td>cingulate sulcus</td>
<td>56 -36 38</td>
<td>3.6339</td>
</tr>
<tr>
<td></td>
<td>posterior cingulate sulcus</td>
<td>62 -30 32</td>
<td>4.5488</td>
</tr>
<tr>
<td>R cingulate</td>
<td>intraparietal sulcus</td>
<td>64 -30 46</td>
<td>5.7406</td>
</tr>
<tr>
<td></td>
<td>cingulate sulcus</td>
<td>48 -34 32</td>
<td>4.3233</td>
</tr>
<tr>
<td></td>
<td>cingulate sulcus</td>
<td>56 -36 38</td>
<td>3.6339</td>
</tr>
<tr>
<td></td>
<td>posterior cingulate sulcus</td>
<td>62 -30 32</td>
<td>4.5488</td>
</tr>
<tr>
<td>R insula</td>
<td>anterior insula</td>
<td>12 -38 58</td>
<td>4.2889</td>
</tr>
<tr>
<td>R subcortical</td>
<td>putamen</td>
<td>38 6 4</td>
<td>5.6508</td>
</tr>
<tr>
<td></td>
<td>amygdala</td>
<td>20 2 4</td>
<td>3.9817</td>
</tr>
<tr>
<td></td>
<td>para-hippocampus</td>
<td>26 -8 -14</td>
<td>4.0395</td>
</tr>
</tbody>
</table>
The modality measure had a significant bilateral effect in the anterior insula and in the lower section of the central sulcus and the posterior superior frontal gyrus (4-3.A,C). Bilateral activation was also observed in the putamen. Right lateralized activation was observed in the supra-marginal gyrus and intraparietal sulcus as well as in the medial and posterior cingulate (4-3.D) and the amygdala.

Peaks of positive difference between the TOM measures and the modal measure (belief + intentions> modals) ($P < 0.005$ uncorrected, cluster threshold =20) are reported in table (4.7).

<table>
<thead>
<tr>
<th>Region</th>
<th>Location</th>
<th>MNI coordinates ($x,y,z$)</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>L temporal</td>
<td>middle temporal gyrus</td>
<td>-60 -14 -26</td>
<td>4.3785</td>
</tr>
<tr>
<td>L frontal</td>
<td>anterior superior frontal sulcus</td>
<td>-12 60 36</td>
<td>7.0077</td>
</tr>
<tr>
<td>L frontal</td>
<td>middle frontal gyrus</td>
<td>-28 28 54</td>
<td>5.8215</td>
</tr>
<tr>
<td>L frontal</td>
<td>anterior superior frontal gyrus</td>
<td>-36 30 50</td>
<td>3.9145</td>
</tr>
<tr>
<td>L parietal</td>
<td>angular gyrus</td>
<td>-46 -60 30</td>
<td>6.9011</td>
</tr>
<tr>
<td>L parietal</td>
<td>precuneus</td>
<td>-44 -80 22</td>
<td>3.1846</td>
</tr>
<tr>
<td>R temporal</td>
<td>middle temporal gyrus</td>
<td>62 -8 -22</td>
<td>4.2649</td>
</tr>
<tr>
<td>R frontal</td>
<td>anterior inferior frontal gyrus</td>
<td>48 -2 -36</td>
<td>4.1425</td>
</tr>
<tr>
<td>R frontal</td>
<td>temporal pole</td>
<td>44 6 -38</td>
<td>3.473</td>
</tr>
<tr>
<td>R frontal</td>
<td>superior frontal gyrus</td>
<td>50 10 -28</td>
<td>3.2039</td>
</tr>
<tr>
<td>R frontal</td>
<td>anterior superior frontal gyrus</td>
<td>10 36 54</td>
<td>4.0303</td>
</tr>
<tr>
<td>R frontal</td>
<td></td>
<td>14 24 60</td>
<td>4.6341</td>
</tr>
<tr>
<td>R parietal</td>
<td>middle frontal gyrus</td>
<td>42 16 44</td>
<td>4.3742</td>
</tr>
<tr>
<td>R parietal</td>
<td>angular gyrus</td>
<td>54 -64 36</td>
<td>5.8982</td>
</tr>
<tr>
<td>R parietal</td>
<td>precuneus</td>
<td>2 -50 34</td>
<td>3.7254</td>
</tr>
</tbody>
</table>
The contrast between TOM and modality had a bilateral effect in the angular gyrus (TPJ), anterior temporal lobe and middle frontal gyrus (4-3.A,C) as well as the precuneus (4-3.B,D). Right lateralized effects were observed in the temporal pole and superior frontal gyrus (dorsal and anterior).
Figure 4-3: A: Lateral view of positive correlation with the modal measure (yellow clusters in the insula and central sulcus) and with the contrast TOM-modal (red clusters in the TPJ, anterior temporal lobe and superior frontal cortex). B: Medial view of positive correlation with the contrast TOM-modal (red cluster in the precuneus) in the left hemisphere. C: Lateral view of positive correlation with the modal measure (yellow clusters in the insula, central sulcus and inferior parietal lobe) and with the contrast TOM-modal (red clusters in the TPJ, temporal pole and dorsal frontal cortex) in the right hemisphere. D: Medial view of positive correlation with the modal measure (yellow clusters in the anterior cingulate and medial superior parietal) and with the contrast TOM-modal (red clusters in the precuneus and anterior prefrontal cortex) in the right hemisphere. All data presented on a two dimensional representation of an 'inflated' cortex. The group t map was thresholded at $p < 0.005$, uncorrected, cluster threshold $\leq 20$. 
4.7 Discussion

4.7.1 Identification of the Theory of Mind network

Our first goal was to find out whether our novel methodology can replicate the general finding in the literature regarding the TOM network. The sentence level model, using a parametric modulation in the spirit of Saxe and Powell (2006) was conceived for this purpose. In (4-1) we plotted the peak coordinates of the main clusters from a recent TOM meta-analysis (Spreng et al. 2008 in blue) in order to facilitate this comparison.

The activations observed here overlap with the meta-analysis peaks in a number of key areas: the bilateral TPJ, the medial anterior (and ventral) prefrontal cortex, the ventral inferior frontal cortex and the precuneus as well as the left anterior middle temporal gyrus. These regions are most often considered as the core TOM network (cf. Buckner et al. 2008). In fact, Saxe and Powell selected a subset of these regions (bilateral TPJ, precuneus and the prefrontal regions) as ROIs for their main experiment, putting aside all other regions found active by the localizer. An additional overlap can be observed in the left fusiform, a region not often discussed in the context of TOM. The only regions observed in our experiment not found in the meta-analysis are the bilateral middle frontal gyrus and the right intraparietal sulcus (though for the latter see Hamilton and Grafton 2006).

There are also a number of regions described by the meta-analysis that were not observed here: The amygdala, which has been considered by some authors a central component of TOM (Baron-Cohen et al., 2000), the dorsal cingulate, the insula, the temporal pole, and the right occipital pole. Interestingly, these regions were found active by some of the more specific contrasts discussed below. We will return to this difference in the context of these more specific findings. Foreshadowing this discussion, we will propose that activation in these regions might support specific aspects of the canonical TOM (false belief) task that are not part of the narrow TOM network as define here. Since our study did not use an explicit false belief judgment task, these additional regions were not observed in the general contrast.

Overall, we find that our results replicate remarkably the findings in the literature.
This is by itself an important discovery since we used a naturalistic stimuli and task (not specifically designed to probe TOM) and a non-standard design. The fact that such design produces such highly interpretable results serves as a proof of concept for this novel approach, and opens the door for future experimentation along these lines. As will be discussed below, this naturalistic, passive task also allows us to distinguish between regions that are part of the narrow TOM network and regions that relate to other aspects of the canonical false belief task.

4.7.2 Intentions and beliefs

As can be easily gleaned from (4-2), activation associated with the belief state measure (in red) largely overlaps with the TOM activation observed in the sentential model (reproduced in light blue for the purpose of comparison). In addition, this measure also correlated with activity in the right middle temporal lobe, the bilateral temporal pole/anterior temporal lobe and in the right occipital lobe. Fusiform activation was observed on the right (contrary to left lateralization observed in the sentence model) and no activation was observed in the ventral inferior frontal gyrus. This overlap is not surprising since the two measures are similar (though the quantification is done in different granularities) and due to the BOLD effective time window they are made even more similar.

The intention measure had a much weaker effect (yellow/green) which implicated, to a large degree, a subset of the regions seen active in the belief contrast (overlap coded in green). Activation was seen in the bilateral TPJ, the right temporal pole, and the right anterior prefrontal cortex. This overall weaker effect is very probably a reflection of differences in power. The belief state parameter had 113 instances in the design while the intention condition had only 51. This important difference, in addition to the general caution when interpreting differences between contrasts does not allow us to infer that regions that do not show an effect of intentionality are only sensitive to belief state. Such an inference can be done by directly contrasting the two parameters (in purple). The direct contrast reveals differences (belief more than intention) in the medial prefrontal cortex, anterior angular gyrus, posterior/dorsal
cingulate and the right middle temporal region. The opposite contrast produced an effect in the left hippocampus.

The angular gyrus (in the temporal parietal junction or TPJ) appears to play a crucial in false belief tasks. Samson et al. (2004) describe three patients with lesions to the left TPJ. These patients exhibited a specific impairment on both verbal and non-verbal versions of the false belief task. Importantly, the same patients were successful on control tasks, which were similar apart from the TOM component. In a follow up experiment, Apperly et al. (2004) demonstrated that patients with frontal lesions that also exhibit TOM deficits failed on the same control tasks suggesting that the role of these regions in the false belief task is of a more general nature (the authors describe it as an executive function impairment).

In the context of the study here the fact that left TPJ damage produced TOM specific deficits is in line with the left lateralization of the TPJ activation found here for all three measures of TOM. Saxe and Wexler (2005) argue for a right lateralization of the TOM specific TPJ activation based on their study. More research is required to resolve this question.

The activation seen in the temporal pole region is in line with the general findings in the fMRI literature (Spreng et al., 2008). The temporal pole is connected (in the macaque) to both posterior visual and auditory regions as well as to the amygdala and the orbital frontal cortex (Olson et al., 2007). Given its location and connectivity pattern, it is not surprising that it has been considered an association cortex. In particular, given its connection to the amygdala, it has been often implicated in social and emotional processes. Olson et al. (2007), reviewing the literature on the topic, conclude that the right temporal pole serves as an interface between complex auditory and visual input (such as faces, narratives or complex sounds) and emotional responses. The left temporal pole also serves as a similar interface, but with semantic memory rather than emotion. Interestingly, our intention measure had a right lateralized effect in this region.

Shaw et al. (2004, 2007) have demonstrated that the temporal pole is not necessary for success on TOM tasks supporting the conclusion that its implication in such tasks...
by imaging tasks (such as the one here) is related to an auxiliary aspect of the task, such as emotional response. It is also of possible significance that only the word measures but not the sentence measure produced an effect in this region. The anterior temporal lobe has been associated with semantic interpretation of single words (Scott et al., 2003) and so one possible explanation for this difference could be that while the TOM/intentional words evoke an emotional response by the nature of their connotations in long term semantic memory, the sentences as a whole were not of particularly high emotional valance. The opposite pattern was observed in the orbital inferior frontal cortex (BA 47). The sentential measure had a bilateral (though left lateralized) effect in this region while neither of the word level measures produced a significant effect. We will return to this distinction in the general discussion when we will consider the role of this region across the different dimensions of sentence processing.

Both the sentential measure and the belief measure produced large effects in the bilateral precuneus region. The intentional measure had no significant effect in that region. Bearing in mind the issues discussed above regarding the interpretation of absence of activation, we find this dissociation potentially important. In 2.8.2 we have discussed the role of the precuneus in a broad domain of cognitive domains (TOM, episodic memory, navigation, syntax, cf. also Cavanna and Trimble 2006; Spreng et al. 2008). We have suggested that this region might be engaged in the representation or processing of hierarchical structure (required by all these different domains). This capacity, though crucial for complex representations such as false beliefs, is not specific to TOM. It is arguably the case that the representation of intentions (contrary to that of beliefs) does not involve the same level of complexity and hence does not generate significant activation in this region. Alternatively, this could be stated as a syntactic difference. Belief operators are most often associated with tensed complementation ('John believes that Mary ate the chocolate') while intentional verbs (in English) usually take non-finite complements ('John wants Mary to eat the chocolate').

The middle temporal lobe (in particular on the right) exhibits a similar dissociation between beliefs and intentions. While this region has been implicated by TOM
tasks (Spreng et al., 2008) the exact role of this region (that could shed light on the origin of the difference observed here) has not been explored in the literature. A possible linguistic explanation for the difference is the richer morphology and syntax (tensed complementation) associated with belief state verbs. Indefrey and Levelt (2004), in a meta-analysis of word processing, associate the anterior middle temporal region with lemma, or syntactic frame, processing. However, these authors report a more anterior and left lateralized pattern while here the pattern is the inverse.

It is also possible that the difference observed in this region reflects the conceptual difference between intentions and beliefs. Vigneau et al. (2006), based on the results of a large scale meta-analysis of language processing in the left hemisphere, describe this region as an amodal conceptual region (together with the angular gyrus) and associate it with a semantic network which includes also the temporal pole and orbital inferior frontal cortex.

The medial frontal cortex and anterior cingulate also appear to be preferentially active for belief compared to intention. Bird et al. (2004) report a case study of a patient with an extensive medial frontal lesion that seems to include the regions found here. The authors did not find the patient to be impaired on a number of TOM related tasks (while executive functions were impaired). Interestingly, none of the TOM tasks involved false beliefs. This is important, since one actual proposal is that the medial prefrontal regions play a role in inhibition of the subject’s own belief state in favor of the belief state of someone else (Leslie et al., 2004). The processing of intention does not necessarily require such inhibition. In linguistic terms (cf. de Villiers 2007) this could be described as the difference between realis and irealis complementation. Only realis complementation is in possible conflict with the state of belief of the subject and hence requires inhibition. The finding here suggests that such inhibition (or enhanced control) is engaged in general for the processing of belief states and not only in the context of a false belief task.

The only region to exhibit more activation for intentions compared to beliefs is the left hippocampus and adjacent regions. Spreng et al.’s (2008) meta-analysis also reports an activation peak in the left hippocampus. Importantly, about half of the
papers that were entered into the meta-analysis probed the representation of others’ intentional state (rather than belief state). Unfortunately, the authors do not report which (or how many) of the papers reported an effect in the left hippocampus. The hippocampus is known to support the representation of cognitive maps. These maps have been proposed to support the representation of goal oriented action or intentions. It is possible that (as is the case in other aspects of TOM), the same region that is used to support the representation of our own goals and intentions is also activated in response to goals or intention of another.

4.7.3 TOM and modality

The main observation from the comparison between TOM and other modals (4.6 - 4.7 and 4-3.A-D) is that activation in the major network of regions implicated by TOM here and elsewhere (Spreng et al., 2008), including the bilateral TPJ, precuneus, middle and anterior temporal and medial prefrontal regions, is not due to the modal component of TOM. Earlier imaging work on TOM did not attempt to tease apart TOM from modality (most often the control condition which served as a contrast to the TOM condition contained less if any modal components, though cf. Fiddick et al. 2005).

A second important observation is that the modal measure produced a significant effect in the right amygdala. The amygdala has often been implicated in TOM processing (Baron-Cohen et al., 2000), in view of its role in mediating social behavior and emotions. At the same time Shaw et al. (2004), studying patients with lesions to the amygdala, did not find an associated TOM deficit, arguing against a necessary role for these structures in TOM. In our experiment, none of the TOM measures produced amygdala activation. It is possible that we lacked power or that the amygdala is engaged preferentially in TOM tasks that require an explicit decision or an inference procedure (neither one required by our passive task). Neither of these two lines of reasoning predict the amygdala activation by the modal measure. The activation peak in the right amygdala for the modal measure ([22, 0, -14]) is remarkably close to the peak proposed by Spreng et al.’s (2008) meta-analysis of TOM ([22, 1, -12])
strongly suggesting that this activation is related to the one observed for TOM in the literature. What could explain this activation? We would like to propose, tentatively, that the amygdala’s activation in TOM tasks is mediated through its role in the processing of modality. Specifically, we note that modal propositions often engender uncertainty (‘John might come’) and that the amygdala plays a role in the processing of uncertainty (Hsu et al., 2005). We thus propose that TOM related activation in the amygdala is (at least in certain cases) related to its more general role in the representation or processing of uncertainty.

Finally, the modal measure produced significant activation in the bilateral insula. One possible explanation for this activation is that most modal terms we coded for are auxiliaries (such as might, would, will) which can be unusual phonologically. The insula has been implicated in speech processing (for perception cf. Bamiou et al. 2003 and production cf. Ackermann and Riecker 2004). This line of explanation might also be relevant for the activation observed bilaterally in the putamen.
Chapter 5

Non-local dependencies

Syntagmatic relations – such as the ones that hold between a verb and its direct object; a subject and a verb phrase; or a preposition and its complement – are most often local in nature. However, one hallmark of natural language is the existence of non-local relations as well. It is possible to divide non-local dependencies into two sub-groups. On the one hand, there are anaphoric relations between a (possibly null) pronominal term and its antecedent. On the other hand, are dislocations, where a linguistic unit appears in a position which is distinct from the position where it is intuitively interpreted (marked by an underscore in 33). In this chapter we will limit ourselves to the study of dislocation:

The exact definition of locality depends on the syntactic framework one assumes. In a constituent grammar, as the one broadly assumed within generative grammar and re-write rule grammars (such as the context free grammar used by Roark’s parser), locality is often reduced to sisterhood relation. However, in dependency grammars, which make no reference to constituents, locality needs to be defined by some other metric, such as the nature of the intervening material. However, it is most often the case that both metrics are used (Nivre, 2005). In generative grammar, specifier-head relations are one example of a local relation which is not strictly a sisterhood relation.

Ross (1967); Langacker (1969) have proposed that pronominal reference should also be treated via a transformation. This idea has resurfaced in more recent discussion of particular cases of pronominalization such as pseudo-clefts (Percus, 1997). However, it has become standard to treat the two phenomena as distinct. cf. Callahan (2008) for discussion of the two types of non-local dependencies in the context of electrophysiology.
Every boy hoped that the teacher will post his picture on the class website.

The apple, Mary thought Bill has offered to Peter.

5.1 The neurolinguistics of non-local dependencies

The analyses reported here were designed with three questions in mind, each on a different level of the (Marr and Poggio, 1976) hierarchy:

- A computational-level question: The A /A-bar distinction.
- An algorithmic-level question: Maintenance versus cue based retrieval strategy in the processing of A-bar dependencies.
- An implementation-level question: Brain network supporting non-local dependency processing.

We will discuss the three questions in turn, briefly reviewing the relevant literature. We will then describe the approach taken here to address these questions and the relevant results.

5.1.1 The A/A-bar distinction

In English, the term ‘A-movement’ groups together a number of (putative) non-local relations, where the surface position of the dislocated element is a either the subject position (passives 34a, unaccusatives 34b, raising to subject 34c) or, under certain analyses (Postal, 1974), the object position (raising to object 34d), that is, an argumental position:

\[3\]

Baltin (2001) defines an argumental position as “a position in which an element can be base generated and bear a crucial semantic role with respect to the main predicate of a clause” (p. 226). In other languages, A-movement has been argued to include also other dislocations such as scrambling and object shift. Since the late 1980’s many researchers within generative grammar have assumed that all non-expletive subjects in English are A-dislocated to that position (the so called VP internal
(34) A-movement:
   a. Bill$_i$ was kicked $\rightarrow_i$ by the wicked boy
   b. The picture$_i$ dropped $\rightarrow_i$ off the table
   c. The cat$_i$ seems to Mary $\rightarrow_i$ to be asleep
   d. The detective figured Bill$_i$ out $\rightarrow_i$ to be the thief

The term 'A-bar movement' groups together dislocations in which the surface position of the dislocated element is not an argumental position (e.g. topicalization: 33, Wh-movement: 35a, relative clauses: 35b)

(35) A-bar movement:
   a. Which cake$_i$ did the office manager eat $\rightarrow_i$?
   b. Paul read the book$_i$ that Mary recommended $\rightarrow_i$ to him

While both types of constructions exhibit a non-local relation between (usually) a predicate and an argument, they differ in a number of important ways. First, A-bar dependencies are potentially unbounded and can extend cross multiple clauses (36). This is not the case for A-dependencies, which cannot span across a tensed clause boundary (The star in 37 indicates unacceptability):

(36) Which carrot$_i$ did Paul believe that Norman said that the donkey ate $\rightarrow_i$?

(37) *Peter$_i$ seemed that $\rightarrow_i$ liked Dora

Binding (or the generalizations governing coreference between noun phrases in a sentence) also distinguishes the two construction types. In A-bar constructions the displaced element is subject to the binding conditions holding on the base position (in 38 the displaced embedded subject cannot corefer with the c-commanding pronoun). This is not the case for A-dependencies (coreference is allowed between the experiencer 'each other' and the raised noun phrase in 39):

(38) *John$_i$, who he$_i$ thought $\rightarrow_i$ would win ... (Baltin, 2001)

subject hypothesis, Koopman and Sportiche 1991. Here we assume a more narrow, or descriptive, use of the term that covers only 'non-canonical' subject configurations.
(39) They see each other to be polite (Baltin, 2001)

Bresnan (1978) notes another important distinction between the two families of constructions. The conditions governing A-movement are functional (defined by terms such as ‘object of’ or ‘subject of’) while A-bar transformations are defined structurally. Relatedly, A-movement is often lexically conditioned and, in cases such as the passive, has a morphological reflex on the verb.

Despite the above mentioned differences, in the transition within transformational generative grammar from multiple specific transformations (Chomsky, 1965) to a single, general notion of move (Chomsky and Lasnik, 1977), Chomsky and others have proposed that both types of dislocations are mediated by that general syntactic transformational operation. This assumption has been maintained in later frameworks (namely Government and Binding, Chomsky 1981, and Minimalism, Chomsky 1994).

Bresnan (1978), building on earlier work within Relational Grammar (Perlmutter and Postal, 1977), arrived at a different conclusion. In the framework that emerged from that paper (Lexical Functional Grammar, Bresnan and Kaplan (1982)) only A-bar dependencies are assumed to involve ‘true’ syntactic transformations. A-dependencies are relegated to the lexicon, where (often verb specific) lexical rules govern the mapping from functional structure to surface structure. The subject of a passive sentence is not related transformationally to the syntactic object position. Instead, the functional object role is mapped, by a passive lexical rule, to the syntactic subject position. This ‘dual mechanism’ framework (in contrast to the single syntactic mechanism proposed by Chomsky) has been adopted by other lexicalized formalisms such as HPSG (Pollard and Sag, 1994) and Role and Reference Grammar (Van Valin, 1993).

The competence level debate between a single and a dual mechanism, has made its way to neurolinguistics and in particular to the debate regarding the nature of Aphasia. Grodzinsky and colleagues (Grodzinsky et al., 1991; Grodzinsky, 2001) have found that agrammatic aphasics exhibit comprehension deficits on both types of dependencies and argued that this is evidence for a shared core transformational mechanism. While no work known to us has tried to argue for the opposite grammat-
ical view, much recent work has argued against the categorical nature of the deficit pattern described by these authors (cf. commentary papers in Grodzinsky 2001 and Gahl et al. 2003).

Neuroimaging has a potential of to shed light on the single/dual mechanism debate since neural overlap in the processing of the two construction types (or the absence of such overlap) can be used as relevant evidence. However, direct comparison, using fMRI, between A and A-bar constructions is complicated by the fact that apart from the putative shared mechanism, many other aspects of the constructions differ and so the results of a standard subtraction analysis would be hard to interpret.

The alternative, of comparing each of the constructions to a different baseline and then comparing the difference maps, is subject to the opposite problem. Both constructions could differ from a baseline on similar dimensions other than the presence of a dependency (e.g. canonicity, animacy order and information structure) causing overlapping activation that is not due to a shared transformational core. In effect, we are not aware of any published attempt (at least for English) to contrast or compare the two types of displacement.

In the experiment here we tried to address this question by adding both an A and an A-bar regressor to the model. We did not compare one type of A-bar construction (for example, Wh-question) to one type of A construction (for example, passive) but instead collapsed across multiple types of constructions from each family (wh-questions, relative clauses and clefts for A-bar and passive, raising, control and unaccusatives for A). This cross construction comparison provides for an implicit orthogonalization of the variable of interest (the dependency) from other construction specific properties. Secondly, these regressors were entered to a model that contained many other linguistic variables (mostly discussed in the preceding chapters) such as surprisal, ambiguity and syntactic complexity. The other regressors served to control

\footnote{Control has been traditionally not considered a case of A-movement but rather a case of phonologically null anaphoric relation. However, recent theoretical work Hornstein (1999) and experimental work (Walenski, 2002) have suggested the possibility of a unified account. In the work here we opted to group the two together.}
for possible confounds between the effect of the dependency itself and the effect of other, possibly related aspects of the linguistic structure.

5.1.2 Maintenance and retrieval of A-bar dependencies

A-bar dependencies have been often used to study verbal working memory (Caplan and Waters, 1999) since they exhibit all the elements of a standard working memory task, with explicit encoding (the marked antecedent), a maintenance period and a retrieval probe (at the gap). The classic model of working memory (Baddeley, 1976) contains a phonological buffer that serves to actively maintain the encoded element in working memory until it is retrieved. Gibson's syntactic implementation of this model (Gibson, 1998, 2000) followed this assumption in associating a cognitive cost with the maintenance of the antecedent until it is reactivated (or retrieved) at the gap.

Recently, an alternative general view of working memory (Cowan, 1995; McElree, 2006) and specifically syntactic working memory (McElree et al., 2003; Van Dyke and McElree, 2006; Lewis and Vasishth, 2005) has advanced a model with no explicit maintenance component. Instead, encoding is viewed as temporary boosting of a long term memory trace (no specific working memory representation). After encoding, the activation of that memory trace starts to decay. Cost is incurred only at retrieval, where the distance from encoding and number of competitors come into play. Greater decay (caused by longer distance) or larger number of competitors require more retrieval effort.

Though both models expect a cognitive effect at retrieval, only Gibson's model expects an effect associated with maintenance per-se. Two sources of evidence for maintenance-related cost are ERP data (King and Kutas, 1995; Phillips et al., 2005), that show a sustained anterior left negativity during the maintenance period and reading time data. For example, Gibson and colleagues (Chen et al., 2005) found that the reading time for linguistic material (in bold in 40) in the presence of an open dependency was significantly longer than the reading time for the same material without such dependency (41):
The claim, which the cop who the mobster attacked ignored might have affected the jury.

The claim alleging that the cop who the mobster attacked ignored the informant might have affected the jury.

Lewis et al. (2006), reviewing the evidence for a maintenance cost propose that in a model without an explicit maintenance component (as the one advanced by these authors) apparent maintenance effects as in (40) could be due to increased interference during the processing of the bold-faced verb attacked since the A-bar displaced element is an active competitor to the cop. This alternative view can be potentially distinguished from the maintenance view in at least two ways. First, the maintenance view predicts constant added cost throughout the maintenance period while the interference account predicts added cost only at the verb (when interference takes place). Secondly, the interference account predicts the same brain network that engages in retrieval to simply be taxed to a greater extent during maintenance, while the explicit maintenance perspective expects (potentially) different regions to be engaged in retrieval and maintenance. Neither ERP or behavioral data can speak to the second point.

Functional MRI could potentially distinguish, spatially, maintenance and retrieval-based activation (cf. Zarahn et al. 2006). However, all work on the topic to date has not tried to distinguish the two components, taking activation over the entire sentence as the dependent measure. Certain authors (Fiebach et al., 2002b) have attempted to probe maintenance cost by manipulating the length of dependency. However, as has been briefly mentioned earlier, both approaches assume that extended dependency length would increase retrieval cost and since the dependent measure integrates activation across the entire sentence, these results are still ambiguous. In the experiment here, we modeled both retrieval (as parametric modulators) and maintenance (as a separate block condition, where each block spanned the maintenance region). Using this approach we could assess, simultaneously, the effects of maintenance and retrieval.
5.1.3 Brain networks supporting non-local dependencies

The final, implementational-level, question regards the brain regions supporting the processing of non-local dependencies\(^5\). Lesion studies (Grodzinsky et al., 1991) as well as early imaging experiments (Just et al., 1996; Stromswold et al., 1996) implicated Broca's region (BA 44 and 45 in the left inferior frontal gyrus) in the processing of syntax more generally, and long distance dependencies more specifically (see also Grodzinsky 2001). As a result, much of the more recent literature has focused on this region, often using an ROI approach (Santi and Grodzinsky, 2007; January et al., 2008) or otherwise downplaying significant activation found in other regions (cf. Shetreet et al. 2007 for one such example). However, a review of the literature strongly suggests that other regions might play a significant role in the processing of long distance dependencies and that Broca's region itself might not be as central to that aspect of linguistic processing as has been previously assumed.

Dronkers et al. (2004) in a large scale lesion study of over 60 left hemisphere lesioned patients, did not find Broca's area to be crucial for the comprehension of complex syntactic structure, which contained a variety of non-local dependencies. Instead, BA 47 (the orbital part of the inferior frontal gyrus) as well as the left temporal parietal junction region emerged as crucial for the processing of these sentences. Caplan et al. 2003, used the same materials as in the original PET experiment reported in Stromswold et al. (1996) to tests older adults and young adults with less education than the original experiment's undergraduate populations. While the 1996 experiment revealed left inferior frontal gyrus activation, the study on older adults found activation in the left anterior superior frontal gyrus (BA10) and inferior parietal lobe (BA 40). Young, less educated, adults exhibited activation in the superior frontal (BA 6) and medial parietal (BA 7) gyre. Neither group presented significant activation in Broca's area.

Whole brain studies investigating non-local dependencies that do find activation in Broca's area often find relevant activation in other regions as well. For example,\(^5\)

\(^5\)This question has been implicitly addressed in the first two questions since, given our methodology, brain activation patterns will be used to answer the first two questions.
Christensen (2008) used fMRI to investigate the neural correlates of different non-local dependencies in Danish. The Wh-movement condition (A-bar) produced activation in Broca’s region but also in BA 47, the right cerebellum and the middle and posterior superior temporal lobe. The involvement of that last region in non-local dependency processing was also reported in a series of experiments (in Hebrew) by Ben-Shachar and colleagues (Ben-Shachar et al., 2004).

Christensen (echoing a general trend in the literature, cf. Vigneau et al. 2006) describes his findings as an evidence for a distributed brain network supporting syntactic processing. Reviewing a number of recent fMRI experiments, the author suggests the following regions as sub-components of that network: The left frontal gyrus including Brocas area and its vicinity, the left posterior superior temporal gyrus including Wernickes area and its vicinity (the Temporal Occipital Parietal Junction), left anterior temporal superior gyrus and their right-hemisphere homologues, as well as the dorsal prefrontal cortex, the cerebellum, and the anterior cingulate gyrus.

The notion of a distributed network of regions is, of course, just a first step towards a neurolinguistic model. As highlighted by Fiebach et al. 2002b; Christensen 2008 in the context of non-local dependency processing and more generally by recent reviews (Hickok and Poeppel, 2007; Ben Shalom and Poeppel, 2008), the different components of the network need to be functionally characterized. While multiple functional hypotheses regarding the specific role of Broca’s area have been articulated and also put to test (cf. January et al. 2008 for one recent example), the function of other sub-components of this hypothesized network has not received the same attention. This, we believe, is a result primarily of the experimental paradigm. The (common) use of whole sentence measures does not allow to distinguish between separate temporal components of the task.

A second problem is that despite the existence of highly articulated linguistic and psycholinguistic models (cf. Lewis and Vasishth 2005), most fMRI experiments do not attempt to manipulate specific aspects of the phenomenon but instead use more coarse comparisons (usually movement versus lack of movement or object relatives versus subject relatives). Since working memory has been a central organizing theme for the
research on the topic, the only dimension that has been manipulated is the length of
the dependency (Kaan et al., 2000; Fiebach et al., 2005; Santi and Grodzinsky, 2007)
but, as discussed above, due to the coarse time resolution, the interpretation of the
results is complicated.

In the experiment here we attempted to study the role of different components of
the network by quantifying a number of different dimensions simultaneously, namely
the type of dependency (A vs A-bar), the maintenance period versus retrieval period
and the specific effect of competition during retrieval (Gibson, 2000; Van Dyke and
McElree, 2006). The last measure was quantified using Gibson’s Dependency Locality
Theory (DLT). For each gap, the DLT competition cost was equal to the number
of discourse referents (referential, non-pronominal, noun phrases and lexical verbs)
intervening between the antecedent and the gap). This measure has been selected
because DLT has been shown to correlate with reading time (Grodner and Gibson,
2005) and eye tracking measures (Demberg and Keller, 2007, 2008) and because it is
possible to compute manually.

5.2 Non-local dependency measures

In the Word model, non-local dependency processing was modeled using three des-
ignated parametric modulators of the word condition and an additional independent
condition representing maintenance. The three modulators represented A gaps, A-bar
gaps and DLT competition cost.

The A gap modulator had the value 1 for each word introducing an A gap
(unaccusitive verbs, passive verbs, infinitival to in raising and control constructions)
and 0 elsewhere. The A-bar modulator had the value 1 for each word introducing
an A-bar gap (the verb for direct object and subject gaps, the last overt element of
the obligatory complements of the verb for adjunct gaps) and 0 elsewhere.

Ideally, it would have been more accurate to model the gap in a more realistic
temporal position. Cross modal priming at the gap (Nicol et al., 1994) has demon-
strated lexical reactivation in the gap about 600 ms following the onset of the word
preceding the gap. However, given our use of parametric modulators associated with
word onsets, this was not possible. In any case, it is doubtful that a 600ms differ-
ence would make an important difference given the slow nature of the hemodynamic
response (Susan Whitfield-Gabrieli, p.c.).

**The DLT competition measure** was quantified only for A-bar gaps (most A
gaps were too local to produce any meaningful variance along this dimension). For
each A-bar gap (defined the same as for the A-bar measure), the modulator contained
the number of discourse referents (see above) intervening between the antecedent and
the gap.

The traditional view of working memory assumes a constant cost of storage be-
tween the antecedent’s position and the associated gap. Adding a parametric modu-
lator for storage would not be appropriate since it would model storage as an effect on
word processing rather than as a constant cost. Instead, we chose to model storage,
or maintenance, as a block condition\(^6\). The onset of each block in the **maintenance
condition** was determined as the onset of an overtly marked A-bar displaced con-
stituent (e.g. a Wh word) and otherwise the first word that indicates the presence of
an A-bar displacement (e.g. a relative clause *that*). The length of the blocks was not
uniform but reflected the length of the maintenance period, ending at the end of the
word introducing the gap.

The Sentence model included a displacement parametric modulator. The dis-
placement score of a sentence was computed as the number of gaps (A and A-bar)
per words in the sentence.

\(^6\)In fMRI, a block design differs from an event design in that blocks assume that neural activity
go on for a period of time (the length of the block) while events model neural activity as a short
lasting burst of activation. Though most fMRI experiments are exclusively block or event related, a
mixed design (cf. Strange et al. 2005) is also possible. This is the approach taken here.
5.3 Results

We will present the results in the following way. First, we will present the Sentence model displacement measure. Then we will present the Word model results for A-bar gap processing\(^7\) and A-bar maintenance. We will end with the effect of A gaps.

5.3.1 Displacement results (the Sentence model)

Peaks of positive correlation with the displacement measure from the Sentence model \((P < 0.005\) uncorrected, cluster threshold = 20) are reported in table (5.1). The displacement parameter had a positive effect primarily in the left hemisphere (5-1) with significant activation in BA 47 (orbital part of the inferior frontal gyrus) and the superior frontal gyrus in the frontal lobe. Parietal activation was observed in the angular gyrus (in the temporal parietal junction) and the precuneus. Right hemispheric activation (not rendered) was observed in the medial wall (anterior prefrontal, cingulate and precuneus) as well as in the inferior parietal gyrus.

<table>
<thead>
<tr>
<th>Region</th>
<th>Location</th>
<th>MNI coordinates (x,y,z)</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>L frontal</td>
<td>superior frontal gyrus</td>
<td>-22 30 58 38</td>
<td>4.4092</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-36 26 52</td>
<td>3.4837</td>
</tr>
<tr>
<td>L parietal</td>
<td>orbital part</td>
<td>-32 40 -18</td>
<td>4.0853</td>
</tr>
<tr>
<td></td>
<td>angular gyrus</td>
<td>-40 -56 26</td>
<td>4.6158</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-40 -56 36</td>
<td>4.015</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-52 -58 26</td>
<td>3.6593</td>
</tr>
<tr>
<td>L cingulate</td>
<td>precuneus</td>
<td>-6 -66 34</td>
<td>3.8664</td>
</tr>
<tr>
<td></td>
<td>posterior cingulate</td>
<td>-8 -46 12</td>
<td>3.4988</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-10 -46 4</td>
<td>3.1108</td>
</tr>
<tr>
<td>L sub-cortical</td>
<td>cerebellum</td>
<td>-10 -42 -20</td>
<td>4.4736</td>
</tr>
<tr>
<td>R frontal</td>
<td>anterior prefrontal</td>
<td>32 58 0</td>
<td>3.7377</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30 48 2</td>
<td>3.1622</td>
</tr>
<tr>
<td>R parietal</td>
<td>precuneus</td>
<td>18 -58 26</td>
<td>4.3966</td>
</tr>
<tr>
<td></td>
<td>inferior parietal gyrus</td>
<td>62 -50 40</td>
<td>3.5365</td>
</tr>
<tr>
<td></td>
<td></td>
<td>58 -54 34</td>
<td>3.0699</td>
</tr>
<tr>
<td>R cingulate</td>
<td>posterior cingulate</td>
<td>12 -46 4</td>
<td>3.8005</td>
</tr>
</tbody>
</table>

\(^7\)The A-bar gap measure was computed as the combination of the basic gap parameter and the DLT competition cost.
Figure 5-1: A: Lateral view of positive correlation with the sentence model displacement parameter in the left hemisphere (clusters in the TPJ, orbital gyrus, and superior frontal gyrus). B: Medial view of positive correlation with the sentence model displacement parameter in the left hemisphere (clusters in the precuneus and posterior cingulate). All data presented on a two dimensional representation of an 'inflated' cortex. The group t map was thresholded at $p < 0.005$, uncorrected, cluster threshold=20.
5.3.2 A-bar maintenance versus gap related activation

Peaks of positive effect of the A-bar maintenance condition from the Word model (\(P < 0.005\) uncorrected, cluster threshold =20) are reported in table (5.2).

<table>
<thead>
<tr>
<th>Region</th>
<th>Location</th>
<th>MNI coordinates (x,y,z)</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>L temporal</td>
<td>superior temporal sulcus</td>
<td>-52,-26,-10</td>
<td>5.0117</td>
</tr>
<tr>
<td></td>
<td>superior temporal gyrus</td>
<td>-50,-20,-4</td>
<td>4.6738</td>
</tr>
<tr>
<td></td>
<td>occipito-temporal sulcus</td>
<td>-28,-70,0</td>
<td>4.6668</td>
</tr>
<tr>
<td></td>
<td>planum temporale</td>
<td>-58,-34,16</td>
<td>4.7002</td>
</tr>
<tr>
<td>L frontal</td>
<td>precentral sulcus</td>
<td>-24,-12,52</td>
<td>5.2255</td>
</tr>
<tr>
<td>L parietal</td>
<td>superior frontal gyrus</td>
<td>-20,-2,64</td>
<td>4.8557</td>
</tr>
<tr>
<td></td>
<td>supramarginal</td>
<td>-50,-24,42</td>
<td>5.9564</td>
</tr>
<tr>
<td></td>
<td>postcentral sulcus</td>
<td>-42,-34,42</td>
<td>4.1095</td>
</tr>
<tr>
<td></td>
<td>postcentral gyrus</td>
<td>-60,-18,38</td>
<td>4.106</td>
</tr>
<tr>
<td></td>
<td>supramarginal</td>
<td>-58,-24,16</td>
<td>4.4259</td>
</tr>
<tr>
<td></td>
<td>postcentral sulcus</td>
<td>-34,-42,50</td>
<td>4.3456</td>
</tr>
<tr>
<td>L occipital</td>
<td>anterior superior occipital sulcus</td>
<td>-24,-78,24</td>
<td>5.0097</td>
</tr>
<tr>
<td></td>
<td>middle occipital gyrus</td>
<td>-10,-92,14</td>
<td>4.1247</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-16,-90,8</td>
<td>3.99</td>
</tr>
<tr>
<td>R temporal</td>
<td>superior temporal gyrus</td>
<td>62,-12,2</td>
<td>4.6777</td>
</tr>
<tr>
<td>R parietal</td>
<td>superior parietal gyrus</td>
<td>52,-14,2</td>
<td>4.3112</td>
</tr>
<tr>
<td></td>
<td>postcentral gyrus</td>
<td>36,-34,62</td>
<td>3.9582</td>
</tr>
<tr>
<td></td>
<td></td>
<td>32,-42,58</td>
<td>3.7583</td>
</tr>
<tr>
<td></td>
<td>inferior parietal gyrus</td>
<td>52,-22,48</td>
<td>4.8059</td>
</tr>
<tr>
<td></td>
<td>postcentral gyrus</td>
<td>60,-22,50</td>
<td>4.4124</td>
</tr>
<tr>
<td>R occipital</td>
<td>superior occipital sulcus</td>
<td>24,-80,12</td>
<td>5.5048</td>
</tr>
<tr>
<td></td>
<td>calcarine sulcus</td>
<td>24,-72,6</td>
<td>4.627</td>
</tr>
</tbody>
</table>

Maintenance periods were associated with increased activation in the superior temporal lobe, post central gyrus and superior occipital sulcus bilaterally (rendered in yellow in 5-2.A,C). Left lateralized activation was observed in the occipito-temporal sulcus, precentral sulcus and superior frontal sulcus as well as the supramarginal gyrus. Right lateralized activation was observed in the inferior parietal gyrus and the calcarine sulcus.

Peaks of positive correlation with the A-bar gap measure from the Word model (\(P < 0.005\) uncorrected, cluster threshold =20) are reported in table (5.3).

The presence of A-bar gaps had a positive effect primarily in the left hemisphere. Bilateral activation (rendered in green in 5-2) was observed in the the posterior temporal lobe (the Temporal Parietal Junction), the medial prefrontal cortex the insula and the amygdala/hippocampus region. Left hemisphere loci of activation were observed
Table 5.3: The A-bar gap measure results

<table>
<thead>
<tr>
<th>Region</th>
<th>Location</th>
<th>MNI coordinates (x,y,z)</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>L temporal</td>
<td>middle temporal gyrus</td>
<td>-54 -36 -4</td>
<td>4.2434</td>
</tr>
<tr>
<td></td>
<td>middle temporal gyrus</td>
<td>-64 -34 -2</td>
<td>3.7101</td>
</tr>
<tr>
<td></td>
<td>superior temporal sulcus</td>
<td>-50 -46 0</td>
<td>3.39</td>
</tr>
<tr>
<td></td>
<td>middle temporal gyrus</td>
<td>-62 -12 -10</td>
<td>4.1111</td>
</tr>
<tr>
<td></td>
<td>superior temporal sulcus</td>
<td>-56 -20 0</td>
<td>3.1345</td>
</tr>
<tr>
<td></td>
<td>posterior superior temporal sulcus</td>
<td>-50 -54 24</td>
<td>4.1034</td>
</tr>
<tr>
<td></td>
<td>temporale pole</td>
<td>-46 2 -36</td>
<td>4.0212</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-50 14 -32</td>
<td>3.8016</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-46 16 -24</td>
<td>3.2816</td>
</tr>
<tr>
<td>L frontal</td>
<td>inferior frontal sulcus</td>
<td>-40 20 20</td>
<td>4.7522</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-48 18 20</td>
<td>4.1562</td>
</tr>
<tr>
<td></td>
<td>precentral sulcus</td>
<td>-48 4 22</td>
<td>3.3303</td>
</tr>
<tr>
<td></td>
<td>medial superior frontal gyrus</td>
<td>-6 12 64</td>
<td>4.2149</td>
</tr>
<tr>
<td></td>
<td>superior frontal sulcus</td>
<td>-38 -2 44</td>
<td>4.1718</td>
</tr>
<tr>
<td></td>
<td>precentral gyrus</td>
<td>-54 8 42</td>
<td>3.7918</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-50 4 48</td>
<td>3.7732</td>
</tr>
<tr>
<td></td>
<td>orbital inferior frontal</td>
<td>-42 22 -16</td>
<td>4.0116</td>
</tr>
<tr>
<td></td>
<td>medial frontal gyrus (SMA)</td>
<td>0 -16 62</td>
<td>3.1869</td>
</tr>
<tr>
<td>L parietal</td>
<td>dorsal supramarginal</td>
<td>-54 -40 -44</td>
<td>3.7295</td>
</tr>
<tr>
<td>L insula</td>
<td>inferior insula</td>
<td>-38 -12 -14</td>
<td>3.9927</td>
</tr>
<tr>
<td>L sub-cortical</td>
<td>hippocampus</td>
<td>-24 -14 -14</td>
<td>4.2665</td>
</tr>
<tr>
<td></td>
<td>thalamus</td>
<td>-20 -12 4</td>
<td>4.3518</td>
</tr>
<tr>
<td>R temporal</td>
<td>planum temporale</td>
<td>56 -40 18</td>
<td>4.3784</td>
</tr>
<tr>
<td></td>
<td>superior temporal sulcus</td>
<td>60 -40 8</td>
<td>4.1916</td>
</tr>
<tr>
<td>R frontal</td>
<td>medial frontal gyrus (SMA)</td>
<td>6 -26 60</td>
<td>4.4899</td>
</tr>
<tr>
<td></td>
<td>superior frontal gyrus</td>
<td>10 -18 50</td>
<td>3.2508</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6 -10 58</td>
<td>3.7423</td>
</tr>
<tr>
<td>R cingulate</td>
<td>anterior cingulate</td>
<td>16 26 16</td>
<td>6.0422</td>
</tr>
<tr>
<td>R insula</td>
<td></td>
<td>32 6 -12</td>
<td>4.3242</td>
</tr>
<tr>
<td>R sub-cortical</td>
<td>brain stem</td>
<td>4 -16 -16</td>
<td>5.1169</td>
</tr>
<tr>
<td></td>
<td>amygdala</td>
<td>-20 -6 -16</td>
<td>4.801</td>
</tr>
<tr>
<td></td>
<td>cerebellium</td>
<td>16 -46 -20</td>
<td>4.4594</td>
</tr>
</tbody>
</table>

along the superior temporal sulcus from the temporal pole to the temporal parietal junction, throughout the frontal lobe and in the supramarginal gyrus. In the right hemisphere, activation was observed in the anterior cingulate and the cerebellum.
5.3.3 A dependencies

Peaks of positive correlation with the A gap measure from the Word model ($P < 0.005$ uncorrected, cluster threshold $=20$) are reported in table (5.4).

<table>
<thead>
<tr>
<th>Positive correlation with A gaps</th>
<th>Location</th>
<th>MNI coordinates (x,y,z)</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>L parietal</td>
<td>precuneus</td>
<td>-2 36 12</td>
<td>3.656</td>
</tr>
<tr>
<td>L cingulate</td>
<td>anterior cingulate</td>
<td>0 46 2</td>
<td>3.7524</td>
</tr>
<tr>
<td>R temporal</td>
<td>posterior middle temporal gyrus</td>
<td>58 54 2</td>
<td>3.4739</td>
</tr>
<tr>
<td>R frontal</td>
<td>middle frontal gyrus</td>
<td>32 24 48</td>
<td>4.2413</td>
</tr>
<tr>
<td></td>
<td>orbital part</td>
<td>16 28 16</td>
<td>4.0411</td>
</tr>
<tr>
<td>R parietal precentral sulcus</td>
<td>-2 26 42</td>
<td>3.8127</td>
<td></td>
</tr>
<tr>
<td>R cingulate</td>
<td>precuneus</td>
<td>2 54 36</td>
<td>4.5394</td>
</tr>
<tr>
<td>R cingulate dorsal cingulate</td>
<td>2 58 36 4</td>
<td>3.7381</td>
<td></td>
</tr>
<tr>
<td>R cingulate dorsal cingulate</td>
<td>14 36 4</td>
<td>4.7088</td>
<td></td>
</tr>
</tbody>
</table>

Positive correlation with the presence of an A gap in the left hemisphere was limited to the medial wall (anterior cingulate and precuneus) (rendered in blue in 5-2.B). In the right hemisphere activation was observed in the middle temporal gyrus, in the inferior and mid-dorsal frontal lobe, the precuneus and the dorsal cingulate (5-2.C,D).
Figure 5-2: A: Lateral view of positive correlation with the A-bar maintenance measure (yellow clusters in the superior temporal lobe, inferior temporal lobe and postcentral gyrus) and A-bar gap measure (green clusters in middle temporal regions, inferior frontal cortex and inferior parietal) in the left hemisphere. B: Medial view of positive correlation with the A-bar gap measure (green clusters in superior frontal lobe and hippocampus) and A gap measure (blue clusters in anterior cingulate) in the left hemisphere. C: Lateral view of positive correlation with the A-bar maintenance measure (yellow clusters in superior temporal lobe and along the intraparietal sulcus), A-bar gap measure (green clusters in the posterior superior temporal lobe and insula) and A gap measure (blue clusters in posterior middle temporal lobe and dorsal prefrontal cortex) in the right hemisphere. D: Medial view of positive correlation with the A-bar gap measure (green clusters in anterior cingulate and medial frontal gyrus) and A gap measure (blue clusters in the precuneus and cingulate) in the right hemisphere. All data presented on a two dimensional representation of an 'inflated'
5.4 Discussion

5.4.1 The effect of displacement

The left lateralization of the sentence effect of displacement conforms to the general finding in the literature that linguistic, and in particular syntactic processes are left lateralized. The activation in the left temporal parietal junction is also in line with the results of a number of recent fMRI investigations of displacement (Cooke et al., 2002; Constable et al., 2004; Ben-Shachar et al., 2004; Christensen, 2008) and with the lesion data from Dronkers et al. (2004).

The precuneus is known for its role in episodic memory (Cavanna and Trimble, 2006) as well as in verbal working memory (Sabb et al., 2007). Though its role in language processing has gone unobserved (or un-discussed), it has been implicated by modulation of syntactic complexity more generally (Shetreet et al., 2007) and modulation of non-local dependencies more specifically (Caplan et al., 2008).

The orbital region (BA 47) in the left frontal lobe has been repeatedly implicated by syntax (Vigneau et al., 2006) and verbal working memory manipulations (Badre and Wagner, 2007). In effect, Dronkers et al. (2004) found that deficit in the comprehension of syntactic structures containing displacement was associated with lesions to that region, rather than to Broca’s area (BA 44,45) proper.

The activation observed in the superior frontal gyrus was very close to F2p, a region identified by the meta-analysis in Vigneau et al. (2006) as specializing in sentence level processing (this region was also implicated by the sentence level syntactic complexity measure, see 2.10.2). Interestingly, Vigneau and colleagues suggest that F2p constitutes a part of a working memory network, engaged in the processing of complex sentences, together with a region in the posterior superior temporal sulcus (STSp) whose center is adjacent to the TPJ activation cluster identified here.

Activation in the cerebellum has also been reported in the context of non-local dependency processing (Christensen, 2008) but usually right lateralized (reflecting the left lateralization of cortical activation).

In sum, the general pattern emerging from the sentence-level contrast is in agree-
ment with the general literature but also emphasizes the role of regions, such as the precuneus, that have been gone unobserved by earlier studies. The limit of the sentence level model is that it does not allow us to distinguish between different temporal components of the task. Also, by combining A and A-bar dependencies we cannot tell apart their individual contributions. The word-level model results will address these issues.

5.4.2 Maintenance versus retrieval

Both retrieval and maintenance had a distributed effect which was more pronounced in the left hemisphere. Only the gap measure was associated with activation in the TPJ area. This suggests that this region, which has been implicated by multiple experiments, is particularly engaged during retrieval, or reactivation, of the displaced constituent. The retrieval effort required at the gap is reflected also in the observed hippocampus activation. Interestingly, recent functional connectivity work (Kahn et al., 2008) provides evidence for a hippocampal TPJ network.

Another limbic region, the insula, was found to be positively correlated with the A-bar gap measure. Due to the proximity (and connectivity) of the anterior insula to the orbito-ventral frontal cortex, activation in the insula has been often grouped under the term inferior frontal activation (cf. for example Friederici et al. 2000) but it is possible that the insula plays an independent role in language processing. Dronkers et al. (1994) found that lesion to the insula, rather than the inferior frontal gyrus was a better predictor of syntactic comprehension. Stowe et al. (1998) found the left insula to increase its activation with the increase of working memory demands. They suggest (though they don’t temporally dissociate between components) that this region participates in verbal rehearsal. We, however, find this region active at retrieval instead. This alternative view of the role of the insula has independent support from a non linguistic episodic retrieval task (Treyer et al., 2003).

Another region (not activated by the sentence measure) to be activated by A-bar gaps was the inferior frontal gyrus or Broca’s area (BA 45). Badre and Wagner (2007) have proposed that Broca’s area “supports a domain-general selection process that op-
erates post-retrieval to resolve competition among active representations.” (p. 2883).

Recently January et al. 2008 have demonstrated a within-subject co-localization of a stroop task and syntactic ambiguity resolution, arguing for a similar conclusion. The conceptualization of the role of Broca’s region as mediating competition is supported by the finding here as the majority of contemporary models (Gibson, 1998; Lewis and Vasishth, 2005; Van Dyke and McElree, 2006) suggest that reactivation of an antecedent at the gap involves competition between the appropriate antecedent and other active possible targets. The same competition could also be responsible for the activation in the anterior cingulate observed here (and also by Dogil et al. 2002; Röder et al. 2002) since the anterior cingulate has often been implicated in cognitive control and conflict monitoring (Weissman et al., 2005; Fan et al., 2008).

Maintenance was associated with activation in the planum temporale and additional regions in the superior temporal lobe as well as the supramarginal gyrus in the parietal lobe. These regions have been considered part of the phonological network and their activation during maintenance suggests the engagement of a phonological buffer, along the lines of Baddeley (1976, 2000).

Interestingly, the maintenance contrast produced significant activation in the occipito-temporal sulcus, considered part of the visual stream. Though, somewhat surprising, it is worthwhile noting that Cooke et al. (2002) found this region to be implicated in the processing of long distance dependencies (albeit for visually presented sentences). Vigneau and colleagues suggest that this might be a consequence of its role in imagery but it is not clear to us why extended dependencies would produce larger imagery effects. In this respect, the involvement of the occipital lobe in maintenance is even more surprising since it is assumed to be largely dedicated to (low level) visual processing. One possible explanation for this activation could be enhanced attention during maintenance periods (either because of increased effort or because of the anticipation of a gap. Electrophysiological investigations have often identified attention modulated sources in the occipital cortex (cf. Jensen et al. 2007 for a recent review).
5.4.3 A and A-bar gaps

The two gap types produced very different activation patterns. In particular, A gaps were associated primarily with medial activation (precuneus, cingulate) while A-bar gaps activated more lateral regions. Moreover, the A gap activation was more right lateralized. At first view, this lack of overlap appears to support the dual-mechanism architecture (Bresnan, 1978) and argue against the perspective of the transformational perspective. However, the specific pattern of activation suggests an alternative explanation.

From an incremental parsing perspective, A-dependencies differ from A-bar dependencies in that in A-bar dependencies, marking on the antecedent itself, the presence of an overt complementizer, or otherwise impossible NP NP sequences inform the parser regarding the existence of such dependency. In A-movement, it is usually only at the verb (or even later, in the case of unaccusatives) that the parser learns of the existence of such a dependency.

In (4.7.2) we have discussed the role of the precuneus in multiple tasks requiring the manipulation of hierarchical structure, including the syntactic complexity and TOM measures in the experiment here. While the precuneus was implicated in the joint displacement measure at the sentence level, it was only activated by the A gap measure at the word-level model. We suggest that the precuneus is engaged during A-gap processing because the parser is required to reanalyze the syntactic structure which did not include a dependency structure. A lexicalist view of A-movement does not expect any specific syntactic cost but (at most) lexical competition or computation effect. However, the precuneus is not known to support lexical processing.

A second difference is that A-dependencies are usually quite local and involve less intervention between the antecedent and the gap compared to A-bar dependencies. As a consequence, decay of activation of the antecedent and competition from other possible antecedents would be minimal. This, in turn, explains the lack of significant

---

8In this respect it is relevant to note that Callahan (2008), reviewing the ERP and behavioral anaphora processing literature arrives at a similar conclusion, namely that the most significant factor in predicting processing cost at the anaphor is the extent of prior expectation.
activation in the retrieval network (TPJ and hippocampus) as well as in Broca’s area and other regions suggested to support lexical competition and selection.

Despite this alternative explanation for the difference between the two dependency types, the lack of any overlap is still quite surprising. It might also be the case that including in the A measure unaccusatives and control constructions, whose status as A-movement is still under debate, has been a mistake. Further work is required.

5.5 General discussion

At the sentential level the displacement measure produced activation in expected regions given the existing literature with the interesting addition of the precuneus. These results demonstrate, once again, the ability of the current paradigm to replicate results obtained in more conventional designs.

The word-level design results demonstrated that maintenance and retrieval of A-bar dependencies tax separate network arguing against a maintenance free model as the one proposed by Lewis and Vasishth (2005). Broca’s region was implicated by retrieval rather than storage, supporting the view of this region as mediating competition during retrieval (Wagner et al., 2001; January et al., 2008) and arguing against a view of this region as supporting maintenance (Stowe, 2000). To our knowledge this is the first experiment to tease apart these effects in the context of language processing.

Our results suggest that A and A-bar dependencies rely on different neural networks. We have nuanced this result by suggesting that the difference might be due to difference in expectation rather than in the underlying properties of the two dependency types, but further research is required. Specifically, we intend to quantify the predictability of each gap using similar methods to the ones used in chapter 2.

More generally, the ability to quantify the correlates of different aspects of non-local dependencies allows us to break down the description of the distributed network of regions, often observed in the literature to be active in this context, into more specific hypotheses regarding the role of particular regions within that network.
Chapter 6

Towards a cortical model of grammatical competence

6.1 Preliminaries

In this last chapter, we will present a blue print for a cortical model of grammatical processing. This proposal will try to integrate (some of) the results presented in the preceding chapters together with other relevant lesion and functional imaging data (fMRI, MEG and ERP) from the literature, most of which have already been discussed in the relevant chapters. In addition, the model will make reference to cortico-cortical connectivity data (white matter tracts), mostly acquired using Diffused Tensor Imaging (DTI).

The grammatical architecture that will be implemented is not going to be argued, as this would require a much larger format. The model itself is inspired by current theoretical work within generative grammar and most of its components reflect broadly accepted theoretical concepts within that field.

Our model will focus solely on the left hemisphere. We do not doubt that the right hemisphere plays a role in language processing but little is known about its specific role (beyond early acoustic processing, cf. Hickok and Poeppel 2000). Our own data sheds little new light on this issue. The model will also not include sub-cortical regions such as the cerebellum or the basal ganglia, which are known to play
a role in language production and comprehension. This is largely due to the fact that our own data from these structures was not sufficient. Finally it is important to note that this chapter is highly speculative in nature and should be read as such.

6.1.1 Performance vs. competence models

Traditional neurolinguistic models, starting with the classical Wernicke model, have been framed within a *performance* perspective, distinguishing receptive language from language production (as in the Wernicke region/ Broca region dichotomy), visual vs. auditory processing or ‘levels’ or ‘stages’ of processing. In this chapter we chose to characterize the model in *competence* terms. That is, the function of specific regions or sub-networks within the larger model will be specified in terms of a grammatical architecture (cf. Grodzinsky and Friederici 2006). We think of the performance/competence distinction as a difference of perspective rather than as a difference in the domain of study (as it seems to be interpreted sometimes). As a consequence, the competence model is not describing *different* brain regions or brain networks than those described by a performance model but instead conceptualizes the role of largely the *same* regions differently.

It is also not the case that the performance and competence perspectives are completely orthogonal. Since Chomsky’s original division has been primarily a heuristic choice, it is often the case that the study of competence and performance converge. One illustrative example is the notion of ‘late insertion’. Certain theoretical linguists (mostly within the framework known as *distributed morphology*, Halle and Marantz 1993) have espoused a model of grammar where surface lexical items (words) are not atomic primitive units that serve as the building blocks of a larger syntactic structure but are rather the *output* of the syntactic derivation. The syntactic structure is mapped onto phonological forms and so in many cases surface words (for example derivationally complex words or inflected words) represent, in fact, complex syntactic configurations rather than atomic pre-syntactic objects. Interestingly, a very similar concept to that of late insertion has been arrived at by certain neuro-psychologists studying naming in Aphasic patients and healthy controls (A. Caramazza, p.c.).
6.1.2 A Minimalist grammatical architecture

The grammatical architecture that will be assumed here can be described as ‘Minimalist’ (Chomsky, 1994; Marantz, 2005). Given our limited knowledge of the mapping of cognitive structure onto neural tissue as well as the numerous architectural issues which are still under debate within theoretical linguistics, the model here will be highly under-specified. Details of the architecture will be discussed throughout the chapter but a number of central notions will be described here.

Modularity

Phillips (1996) distinguishes two claims regarding modularity: encapsulation and individuation. The two have been historically grouped together in the discussion of language. An encapsulated module (such as in Fodor’s view of input systems; Fodor 1983) does not share information with other modules. For a module to be individuated, it does not need to be encapsulated. Individuation is the ‘claim that there are many distinct modules for many distinct mental functions’ (Phillips 1996,p. 165).

The model proposed here assumes that linguistic behavior is supported by multiple interacting individuated modules 1. As will soon become evident, these modules do not exactly overlap with the traditional semantic/syntax/phonology partition. In that sense, this model distinguishes itself from other recent models, which have kept to this tripartite architecture (most notably, the model proposed by Vigneau et al. 2006, which is the most elaborated neurolinguistic model to date).

The syntactic module

A fundamental assumption within generative grammar is the autonomy of syntax (Chomsky, 1957). Though Chomsky’s original distinction could be interpreted as a methodological one (the autonomy of the study of syntax), it has been implemented architecturally. The syntactic structure and structure-building operations have been taken to exist and operate independently from other aspects of language (such as

---

1some of which might be encapsulated, but this latter question will not be addressed here
meaning or phonological form). Numerous arguments have been produced in favor of this autonomous view. The general form of many of these arguments is the empirical observation that syntactic rules (or transformations) make reference to structural rather than semantic or phonological properties of their triggering context. Data from neuropsychology has also been used to argue for this autonomous view (cf. Dogil et al. 1995).

A second, related, assumption central to generative grammar is the syntacto-centric view of the linguistic sign. Not only is syntax an autonomous level within the language architecture but it is the only level that mediates the relations between the signifier (the phonetic form) and the signified (the conceptual interpretation), to borrow a structuralist terminology. The syntacto-centric principle has often been described using a directionality metaphor (the syntactic structure is built and is then interpreted). This metaphor has been, at times, interpreted (wrongly, in our view) as an essential part of the syntacto-centric principle (the syntax first view)².

The model here implements both of these architectural views. We assume the existence of an autonomous (or individuated) syntactic module. This module is implemented by a network of regions supporting the representation and manipulation of the abstract linguistic hierarchical structure. In addition, our model can be described as syntacto-centric, since both the semantic and phonological components of the models are construed as interpretative modules which interface with the syntactic module but not with each other (cf. Jackendoff 2007 for an example of an alternative architecture).

²F. Ferreira, for example, writes ‘One of the fundamental problems [in the Minimalist Program] is that the model derives a tree starting from all the lexical items and working up to the top-most node, which obviously is difficult to reconcile with left-to-right incremental parsing’ (Ferreira 2005, p. 370). Jackendoff’s (2007) discussion of the main stream generative grammar model as derivational makes the same confound.
The distributed lexicon

The dichotomic distinction between a lexicon and a phrasal syntax is rooted in traditional grammar but has been often espoused both in theoretical linguistics and in psycholinguistics. Under this view the word and the phrase are fundamentally distinct domains. An alternative approach to grammatical architecture replaces the domain based distinction by an operational distinction between storage and computation. Under this view, all structure-building operations are part of the same syntactic module (no lexical specific syntactic or morphological engine). The remanent of the traditional notion of a lexicon is a memorized list.

Inspired by Distributed Morphology (Halle and Marantz, 1993), we will propose a further distinction, within the lexical list, between two modules. The first is the Vocabulary, which contains a list of associations between simplex phonological representations (roots) and non linguistic concepts. Roots have no linguistically relevant internal structure (putting aside the phonological structure). The phonological representation of these roots might be specified in gestural, acoustic or combined code (as has been originally proposed by Wernicke, cf. Hickok and Poeppel 2004). We will have little to say regarding this question.

The second component of the lexicon in the Encyclopedia, which stores certain outputs of the linguistic system, which can then be accessed more rapidly during processing. This system keeps track of the frequency of use of these 'chunks' and any particular idiosyncrasies associated with their use ('world knowledge'), which might not be compositionally computable (as with idioms or novel uses of existing words). There is currently no consensus regarding the criteria determining which linguistic objects can be stored. It could be a categorical distinction (output of certain Phases, see below) or a result of optimization of the grammar (an idea currently being explored at MIT by T. O'Donnell, p.c.), a mix of the two or something else altogether. Once again, we will not argue for a specific proposal here. However, we will assume that the Encyclopedia can store syntactic structures which are larger than surface words. This point will be critical since it will play a role in the conceptualization of syntactic
surprisal effects within the model.

**Semantic interpretation**

In psycholinguistics, the semantic, or interpretative, aspect of language is either treated as a single module or as divided along the lines of lexical vs. compositional meaning. In the model advanced here, this traditional notion of semantics is replaced by a more distributed view.

Much of what is considered lexical semantics is supported by the Encyclopedia and the Vocabulary. The procedure of semantic composition (both in complex word formation and in phrasal syntax) might be individuable as well, however it is very hard to distinguish it from syntactic composition due to the tight coupling between the two (see, for example, Heim and Kratzer 1998). In order to distinguish these two aspects of composition some researchers are studying cases where there is a mismatch between the two, such as in case of complement and aspectual coercion (Pylkkänen, 2008). Our narratives were not designed with the intention of de-coupling syntactic and semantic composition and so currently we cannot address this distinction.

We import from formal semantics the notion of vericonditional meaning, which distinguished propositions (which can be either true or false) from other linguistic units. We take propositions to be the output of the linguistic system that is most relevant to other aspects of human cognition (action, decision making, episodic memory, etc.)

We propose, then, that the representation of vericonditional meaning, or propositional meaning, is another individuated module within the language system. This choice is not a necessary one. It might be the case that events, or situations, are the interface between semantic interpretation and other modules. It might be the case that both propositions and events are interface objects (as seems to be Chomsky’s view when discussing vP vs CP Spell-out; cf. Chomsky 2004). We have not

---

3Semantic research within theoretical linguistics usually does not address the relation between the output of the semantic computation and other cognitive modules. Other frameworks, often associated with Artificial Intelligence research, do make explicit the relation between semantic representation and action.
attempted to address these further possibilities here but find this question to be an exciting future research direction.

Finally, we assume that linguistic behavior makes use of additional modules supporting (among other things) perception and production. However, these have not been central to our work here and will not be addressed in any particular detail. In particular, we will say very little about phonological processing or representations. For a current neuro-cognitive model of phonological perception see Hickok and Poeppel (2007). For a detailed model of speech production see Guenther et al. (2006).

Cognitive modularity and neural segregation

It is important to stress that a modular view does not entail a one-to-one relation between modules and brain regions. It is likely that many modules, in particular within higher cognition, are supported by a network of regions. In fact, this will be largely the case in the model proposed here. Conversely, our current imaging resolution (which is, after normalization, smoothing and group averaging, at best on the scale of about 8 mm) can collapse distinct activations into a single datapoint, masking possibly relevant distinction. Most models are even less specific, with a granularity level of Brodmann’s areas (Brodmann, 1909).

The specificity debate

Putting the neural resolution question aside there is also the issue of cognitive resolution. It is possible that two distinct (individuated) modules make use of overlapping cognitive algorithms and, by extension, overlapping brain tissue. For an hypothetical example take temporal representation. It is possible that the representation of linear order in phonology and in, say, event semantics, makes use of the same mechanism (for example, edge alignment). As a consequence, manipulations of temporal order in phonology and in semantics might produce increased activation in this common network. It appears to us that, in such a hypothetical case, we should not abandon the modular view of phonology and semantics despite the cognitive/neural overlap.

The issue discussed in the last paragraph is particularly important to keep in mind
in cases where the cognitive theory of the individual modules is highly under-specified, making the interpretation of common activation very complicated. This is largely the case in the literature that has attempted to address the relation between language and other cognitive domains (the specificity debate). Often both the experimenters' linguistic assumptions and their assumptions regarding the other domain are too vague to interpret the meaning of an interaction or neural overlap.

In effect, an evolutionary perspective (such as the one advanced by Hauser et al. 2002) stating that linguistic behavior is to a large degree a calque on a pre-existing neuro-cognitive architecture, expects an extensive neural overlap between language and other cognitive faculties. We have no intention to address the specificity question or the evolutionary debate. However, we bring these issues into the discussion here since a number of the regions and even networks that we will make reference to in the model here have been implicated by non-linguistic cognitive research. A full discussion of these overlaps goes beyond the limits of this chapter but we want to make clear that we neither dispute the role of language regions in other cognitive tasks nor do we espouse the non-specific view of linguistic cognition.

**Interfaces**

One major component of the Minimalist architecture is the interface between independent modules. It is proposed that many of the constraints (either universal or language specific) on possible linguistic forms are imposed at these interfaces. This view has been already present in the Government and Binding era (Chomsky, 1981) under the cover of filters (such as the Case filter); however, these filters were assumed to be (syntactic) module-internal. The Minimalist Program represents an attempt to simplify the module internal structure by (among other things) transferring the filtering to the interface between syntax and interpretation (either phonological or semantic). Though grammatical theory is often discussed with a specific directionality (usually syntactic structure is built bottom up and then 'sent' to the interfaces) it is the case that, in a competence model, performance directionality is not at issue and that the interfaces are in principle ‘bi-directional’.
Cyclicity and Phases

Throughout the last 40 years, many researchers have observed cyclic effects in the interpretation of the syntactic structure at the interfaces, that is, evidence for local, or independent, interpretation of sub-structures of the eventual phrase marker. Cyclicity in phonology was discussed already by Chomsky and Halle (1968) and in much subsequent work (cf. for example Bresnan 1971; Kiparsky 1982). Semantic opacity (such as factivity; Kiparsky and Kiparsky 1968), where the interpretation of sub-components of a sentence seems to be independent of the larger sentence has also been observed and discussed. The locality of syntactic transformations (the ‘island’ effects, Ross 1967) has been a central topic in syntactic theory for the last four decades but (for reasons briefly mentioned in the preceding section) has been mostly treated as a syntax internal phenomenon.

Despite the multiple evidence for the locality of the interpretation of the syntactic structures as the interfaces, until the late 1990’s the predominant grammatical model within generative grammar contained a single cycle. In the single cycle ‘T-model’ the entire root sentence is built in the syntax before it is sent, as a whole, to the interpretative modules.

This mismatch between the single cycle model and the abundant evidence for multiple cycles was often dealt with by introducing cyclicity into the interpretation procedure itself. This was done already in SPE (Chomsky and Halle, 1968) and was further developed in frameworks such as Lexical Phonology (Kiparsky, 1982). This solution had the consequence of making the interpretative mechanism quite complex and powerful (and raising the question of redundancy of the syntactic derivation). The consequences of this approach are evident today in Optimality Theory (Prince and Smolensky, 2004), where constraints on phonological form have access to increasingly wide sources of information. Moreover, this module-specific solution could not capture any parallels in cyclic effects across different modules (for example, where phonological opacity correlates with semantic opacity; cf. Bachrach and Wagner 2007).

In the last ten years or so the single cycle model has been largely replaced by a
multiple cycle model (Chomsky, 2001). In this model, the interpretation of syntactic structure takes place at multiple stages of the syntactic derivation and not only at the root. This iterative operation has been named Spell-out and the partial phrase markers which are associated with Spell-out have been termed Phases. There is still no consensus over the correct extensional or intensional definitions of Phases (that is, neither which syntactic structures are Phases nor what makes a syntactic structure a Phase). Similarly, there is no consensus regarding the nature of Spell-out (that is, what interpretative operations take place at every application of the Spell-out function)\textsuperscript{4}.

**Cyclic Spell-out and the lexico-syntactic interface**

Despite the many open questions regarding the notion of Phase, we take the core insight, namely the iterative application of interpretation (Spell-out) to the syntactic structure, to be correct and incorporate it into the model here. One consequence of this view (from both competence and performance perspectives) is that the output of Spell-out of a particular non-root Phase is integrated into the computation at the Spell-out of the Phase containing it.

We follow Marantz (2006) in assuming that to a large extent (at least in English), the pre-theoretic notion word coincides with the ‘first’ application of Spell-out (in a bottom up parlance). Under this perspectives, words are not the output of a separate computation (a lexical syntax, as in Hale and Keyser 1993) or items in a lexical list, but the smallest Phases in a single syntactic computation.

Under this view, the integration of ‘words’ into the larger syntactic structure, which is often viewed (in psycholinguistics as well as by lexicalist approaches of syntax) as an interface between the lexicon and the syntax, is just another case of the integration of the output of a smaller Phase into the output of the Phase containing it (we shall name this operation Phase integration).

This integration operation needs to be distinguished from lexical insertion (or

\textsuperscript{4}In other work (Bachrach and Katzir, To appear), we have advanced a particular proposal regarding some of these issues.
Lexical insertion does have a special status since it involves access to the root Vocabulary which is at the interface of language and the semantic or conceptual storage. In performance terms, lexical insertion is related to lexical access.

Another dimension that relates to the the traditional notion of the lexical-syntactic interface is the Encyclopedia. The Encyclopedia is a repository of Spell-out outputs. These are syntactic objects which have been memorized (according to a criterion which is still unknown), together with their Spell-out interpretation. These memorized syntactic structures can be surface words but also larger constituents (entire verb phrases or noun phrases, frozen expressions, whole idioms, etc.). It is also conceivable that the Encyclopedia includes memorized syntactic frames, that is, outputs of larger Phases where the identity of the smaller Phases contained within them is under-specified (open variables). Certain idioms can exemplify this case. Consider *Give X trouble*. The lexical identity of the first object is left underspecified while the verb, the second object and the syntactic structure (the double object construction) are part of the memorized idiom.

However, the consequences of such a possibility goes well beyond idiomatic expressions (cf. ongoing work by T. O’Donnell at MIT). Work here (chapters 2 & 3) and elsewhere has demonstrated that corpus frequency of certain syntactic structures has an effect on processing (syntactic surprisal and probabilistic ranking of competing parses). In the model proposed here, this effect is relegated to the Encyclopedia as well. We will assume that the Encyclopedia contains a subset of past Spell-out objects (a mental representation that contains the syntactic structure of a Phase as well as its interpretation at the interfaces). Moreover, the Encyclopedia can store syntactic structures in which many (or even all) smaller Phases are under-specified, allowing the brain to keep track of frequency (as well as other features) of these larger structures abstracting away from specific ‘lexical’ choices.
Movement

In the early days of generative grammar, transformations were described as an additional module in the grammar (Chomsky, 1965). Also computationally, transformations required string factorization and context sensitivity that went beyond the limits of the (supposed) context free base. This led researchers to expect to find an inherent cost for transformations in language processing. This expectation was disconfirmed by early psycholinguistic research (Walker et al., 1968). Though, as pointed out by Berwick and Weinberg (1984), this early failure could have been the result of testing the wrong transformational model, a contemporary view of the nature of transformations provides a different perspective.

As was briefly mentioned in chapter 5, generative grammar abandoned particular transformations in favor of a single move alpha transformation already in the early 70's. In the 1990's, with the advent of the Minimalist Program (Chomsky, 1994), the strict segregation between the base component (D-structure) and the transformation component was abandoned. In the new framework basic structure-building operations (Merge) and transformations (Move) were allowed to interleave. However, the two operations where considered fundamentally distinct.

This theoretical distinction between Merge and Move, the remnant of the original distinction between phrase structure rules and transformations, motivated neurolinguistic research to look for a parallel neural distinction. Grodzinsky et al. (1991) proposed that Broca's aphasia is characterized by the loss of the operation Move. Van der Lely and Battle (2003) characterized a sub-type of Specific Language Impairment (grammatical SLI) as a deficit related to the operation Move. This distinction was also tested directly using fMRI by Ben-Shachar et al. (2003, 2004) who claimed that the operation Move can be localized to the the left inferior frontal gyrus (Broca's area).

In recent years, theoretical work within Minimalism has suggested the elimination of the special status of Move, making it into a specific case of Merge (Chomsky, 2004; Starke, 2001), namely Re-Merge (cf. Bachrach and Katzir To appear for discussion
and a specific view of re-Merge). Under this new theoretical view, there is no a-priori expectation to find a neurological dichotomy, in the absence of a theoretical one. It is of course the case that the known features of non-local dependencies (both linguistic and neuro/psycholinguistic) still need to be accounted for. However, the new theoretical perspective suggests that the eventual account will not make reference to a primitive grammatical distinction between structure-building and transformations.

One aspect of grammar which might still be specific to movement is the semantic operation of abstraction. The assumption that the interpretative reflex of movement (or Re-Merge) is an abstraction over the base position of the moved element has been generally accepted within generative grammar (cf. Heim and Kratzer 1998). The converse (that all cases of abstraction are the reflex of movement) is not as widely accepted and researchers differ on issues such as quantifier interpretation, verb valency operations and other domains which appear to require semantic abstraction but do not exhibit other movement-like properties (for example, quantifier raising in English does not induce any word order changes).

In the work reported here we did not quantify directly semantic abstraction. However, we take the sentence level measure of displacement (number of gaps per word in the sentence) to represent (among other things) the extent of movement induced abstraction. We will incorporate this operation in the model to be described here.

6.1.3 Hodology

Hodology (from the greek hodos, road) is the study of the brain’s connectional anatomy (ffytche and Catani, 2005; Catani, 2007), that is, the study of the architecture of the white matter tracts that connect different cerebral structures to each other. In humans, progress in this domain (and in particular in the study of long distance connections) was quite minimal throughout the 20th century since the tracing techniques used for this purpose in other primates were not applicable to humans. Most often, authors extrapolated from primate brains to humans (cf. Pandya et al. 1996) but this method is speculative and is particularly problematic in the case of language since it is unique to humans and since the regions supporting language in the
temporal, parietal and frontal lobes are much more developed in humans compared to non-human primates.

Early imaging techniques were not able to trace these white matter connections but the recent development of Diffusion Tensor Imaging (DTI) tractography has brought about a dramatic change. DTI makes use of water diffusion patterns in white matter to identify the local orientation of white matter fibers (Basser et al., 1994). DTI Tractography makes use of these local patterns to trace white matter tracts between pre-specified regions of interest. While this measure has many limits and is prone to false positives (Catani, 2007), it has permitted previously impossible in-vivo tracing of long distance white matter tracts in humans (as well as other primates). A number of recent papers have confirmed the accuracy of this method using either conventional tracing in primates (Schmahmann et al., 2007) or classical dissection methods in humans (Lawes et al., 2008).

Though, to date, much of the work in DTI tractography has been confirmatory in nature, it has also led to new discoveries and has weighed in on some long standing debates in the literature. In the next section we will review DTI tractographic data regarding the major association fibers that will be implicated in our model. It is important to keep in mind that DTI does not provide data regarding the directionality of the connection, so the use of directional terms should not be taken to entail such directionality.

6.1.4 Major long association fibers serving cortical language regions

The arcuate fasciculus

The arcuate fasciculus (AF) has been traditionally described as a fiber bundle emerging from Broca’s area (BA 44, BA 45), running posteriorly along the sylvian fissure and then looping around and descending into Wernicke’s region in the posterior superior temporal gyrus. The AF has been a central component of Wernicke’s language model, connecting Wernicke’s and Broca’s region. Wernicke himself proposed that lesions to
this fiber bundle are the cause of the conduction aphasia syndrome. Later research has questioned this specific claim but the role of this fasciculus in language processing has not been disputed. Recently, Duffau et al. (2002) have demonstrated that intra-operative direct stimulation of the AF produces language production deficits. Perhaps unsurprisingly, the AF in humans is substantially different than in other primates (Rilling et al., 2008) and so before the advent of DTI little progress has been made in its description beyond Dejerine’s work in the 19th century (Catani et al., 2005).

The use of DTI has proposed a number of significant modifications to the classical view of the AF. Catani et al. (2005) have identified three separate segments within the AF. The first, a direct route from the temporal lobe to the inferior frontal gyrus (similar to the classical view), a second, posterior, section running from the temporal lobe into the inferior parietal lobe, and finally an anterior section running from the inferior parietal lobe to the inferior frontal lobe. Catani and colleagues named the two latter sections the indirect route. Moreover, Catani and colleagues’ DTI tractography data suggests that the AF connects regions also outside the classic language areas; in the frontal lobe, BA47 (part orbitalis), the middle frontal gyrus and the inferior precentral gyrus, and in the temporal lobe the middle temporal gyrus.

Lawes et al. (2008) have confirmed Catani and colleagues’ central findings using a novel DTI tractographic method and classical dissection methods. These authors also provide further specifications regarding the AF. The posterior section of the indirect route connects the posterior middle temporal gyrus with both the supramarginal gyrus and the angular gyrus.

Finally, Rilling et al. (2008) found that the AF connects with a region in the inferior temporal lobe in addition to the middle temporal lobe. In addition, these authors report that that MTG is preferentially connected (via the AF) with BA 45 and 47 (pars triangularis and orbitalis) rather than BA 44 (pars opercular).

Summing up, the AF connects the middle and inferior frontal gyrus and inferior precentral gyrus with temporal and parietal regions. It is comprised of two routes. A direct route connects the IFG directly to the temporal lobe. An indirect route
connects MTG and ITG regions to the supramarginal gyrus, passing through the angular gyrus. The supramarginal and angular gyri are then connected to the frontal lobe via the anterior section of the indirect route.

**The superior longitudinal fasciculus**

The superior longitudinal fasciculus (SLF) connects temporal-parietal regions with the frontal lobe. Dejerine and others have proposed that the AF is part of the SLF and used the terms interchangeably but the study of non-human primate SLF demonstrated that the two are separate bundles. Moreover, the SLF in non-human primates was divided into three sub connections (I, II and III). Makris et al. (2005) using DTI tractography in humans, demonstrated the existence of three pathways paralleling the non-human primate sub connections. SLF I connects the medial superior parietal and precuneus with the superior frontal gyrus (BA 8 and 9). SLF II connects the angular gyrus with the middle frontal gyrus (BA 6 and BA 46), running parallel to the AF (but possibly distinct from it). SLF III connects the supramarginal gyrus with precentral and inferior frontal gyrus and seems to overlap with the anterior part of the indirect route of the AF.

**The uncinate fasciculus**

The uncinate fasciculus (UF) is a fiber bundle connecting the anterior temporal lobe to the ventral frontal lobe and in particular the orbital gyrus and the orbital part of the inferior frontal gyrus (BA 47). In addition, the UF connects BA 47 with the hippocampus and amygdala (Kier et al., 2004).

**The inferior longitudinal fasciculus**

The inferior longitudinal fasciculus (ILF) connects occipital and posterior inferior temporal regions with the temporal pole (the most anterior part of the temporal lobe). The existence of this fiber bundle in humans was the subject of a long standing debate but recent DTI imaging results confirm its existence (Catani et al., 2003). Rushworth et al. (2006) found that the ILF also connects to the angular gyrus.
in the inferior parietal lobe. Mandonnet et al. (2007) suggest that, together with the uncinate fasciculus, the ILF subserves a semantic network engaging posterior temporal and orbito-frontal regions. These authors were not able to produce naming deficits by direct intra-operative stimulation of the ILF. However, they speculate that this might be due to the existence of parallel pathways connecting the frontal and temporal lobes.

**Left superior to inferior frontal gyrus pathway**

Lawes et al. (2008) identified a white matter tract (LSIF), connecting superior and middle frontal gyrus regions to the inferior frontal gyrus and in particular BA 45. This tract will come into play in the model here, since the middle frontal gyrus, traditionally not associated with the brain language network, is taken to be an integral part of the syntactic module proposed here.

**6.1.5 Short association fibers and other cortical connections**

Apart from the long distance connections, a subset of which was discussed in the preceding section, there are also short distance connections between adjacent brain structures. These will not be reviewed in detail here. There are other long distance cortico-cortical connections which were not covered in the preceding section; for example, the connections within the parietal lobe between medial regions (and in particular the precuneus) and lateral sections (the intraparietal sulcus and angular gyrus). Some of these connections have been evidenced by functional connectivity data (correlation of BOLD signal during rest across regions), intra-cranial stimulations and primate studies but have not yet been studied in detail using DTI. We will mention supporting data regarding these connections in the relevant sections of the model description.
6.2 The model

In this section we will describe the details of the model. We will start with an overview of the different components and then turn to more specific discussions.

6.2.1 General outline

The general outline of the model is visualized in (6-1). The different components (with their anatomical distribution and function) are also listed in (6.1):
<table>
<thead>
<tr>
<th>Component</th>
<th>Function</th>
<th>Brain region(s)</th>
<th>Connectivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phonology</td>
<td>Interface of the linguistic system with perception and production.</td>
<td>Superior temporal gyrus (STG) and sulcus, supra-marginal gyrus (SMG), ventral precentral gyrus (VPC), posterior BA 44, fusiform gyrus (reading)</td>
<td>AF, inferior occipito-frontal fasciculus&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Syntax</td>
<td>Representation and manipulation of linguistic hierarchical structure.</td>
<td>Precuneus and middle frontal gyrus (BA 8, 46)</td>
<td>SLF I</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>A list of associations between atomic linguistic signs and non-linguistic concepts.</td>
<td>Middle temporal gyrus (MTG)</td>
<td></td>
</tr>
<tr>
<td>Lexical insertion</td>
<td>The mapping of syntactic terminals onto Vocabulary items.</td>
<td>Posterior superior temporal sulcus (STS)</td>
<td></td>
</tr>
<tr>
<td>Spell-out</td>
<td>Representation of the outputs of the (iterative) interpretation of the syntactic structure.</td>
<td>Ventral inferior frontal gyrus (BA 47-P) and anterior temporal lobe (ATL)</td>
<td>UF, ILF, LSIF</td>
</tr>
<tr>
<td>Encyclopedia</td>
<td>Long term storage of Spell-out objects.</td>
<td>BA 47, anterior temporal lobe (ATL), hippocampus and inferior temporal lobe</td>
<td>UF, ILF</td>
</tr>
<tr>
<td>Pragmatics</td>
<td>Representation of propositional meaning and interface with other cognitive modules.</td>
<td>Anterior BA 47 (BA 47-A)</td>
<td></td>
</tr>
<tr>
<td>Semantics</td>
<td>Computation and evaluation of alternatives.</td>
<td>Medial prefrontal cortex (BA 9)</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>Not discussed here.
Figure 6-1: A graphic summary of the grammatical competence model. The different modules are grouped by colors and full arrows. Major interfaces between modules are marked with white (or black) dotted arrows. See table 6.1 for acronyms.
Auditory/phonological processing (which will not be discussed here in details) takes place in superior temporal gyrus, the supra-marginal gyrus in the inferior parietal lobe, the precentral gyrus in the frontal lobe, and the right cerebellum. In addition, other motor regions (including the basal ganglia and insula) participate in verbal production and possibly perception. For detailed proposals regarding phonological processing see Vigneau et al. (2006); Guenther et al. (2006); Hickok and Poeppel (2007) and references therein.

Syntactic computation (loosely defined) is subserved by a network which includes the precuneus and the angular gyrus in the parietal lobe; the middle frontal gyrus (BA 8), the dorsal prefrontal cortex (BA 46) and the posterior part of BA 47 in the frontal lobe and the anterior temporal lobe.

Conceptual or semantic storage is distributed through the inferior and middle temporal lobe. The Vocabulary (or the association between form and meaning for roots) is implemented in connections between these conceptual regions and phonological forms in the superior temporal sulcus and in connections with orthographic forms in the fusiform gyrus (the word form area). Word formation (or lexical access), which in our grammatical architecture is conceived of as ‘first Spell-out’ (lexical insertion) takes place in the posterior superior temporal sulcus (STS) with BA 45 providing ‘top-down’ control.

The mental representation of the output of Spell-out is subserved by a network composed of BA 47 and the anterior temporal lobe. The Encyclopedia (or stored information regarding Spell-out objects) is implemented within this network and in connections between this network and posterior temporal regions and in particular the posterior MTG. Prosody is implemented in a connection between this frontal-temporal network and more posterior regions in the superior temporal gyrus. Finally, a network composed of the angular and supra-marginal gyri in the inferior parietal lobe, BA 44 and 47 (pars triangularis and orbitalis of the IFG) and posterior MTG subserve the integration of ‘smaller’ Spell-out objects into larger ones.

The mental representation of propositional (or truth conditional) meaning is subserved by the anterior BA 47 and its connection with the hippocampus and the
limbic system more generally. Pragmatic computation has many components and is probably supported by multiple brain regions. Here we will implement explicitly just one aspect of this computation which has received much attention recently in the theoretical literature, namely the computation of alternatives. We propose that alternatives are evaluated by regions within the medial anterior prefrontal cortex. The computation of alternatives requires access to both their syntactic structure and their interpretation (Katzir, 2008). We propose that these medial regions have access to this information through connections to the precuneus (structure) and BA 47 (interpretation).

6.2.2 The core syntax module

We propose that the core system supporting the representation of the hierarchical structure of the sentence, or its syntax, is composed of the precuneus and the anterior and the posterior middle frontal gyrus (BA 46 and 8 respectively). In addition, multiple other regions (BA 47, the angular gyrus, BA 45, etc.) interface with this core system, but these will be discussed separately. SLF I connects the precuneus and the middle frontal gyrus. It has been shown in non-human primates that the precuneus is connected to BA 46 (Cavanna and Trimble, 2006). It is very likely that this network also included sub-cortical structures (and in particular the dorsal striatum; Ullman 2004). However, since our experiment did not produce relevant activation in these structures we will not include them here. From a neural perspective, this network is probably the most novel aspect of the model here, and so also the most speculative.

As we have discussed in 2.8.2, the precuneus has been implicated in a number of tasks or domains that require hierarchical representations (e.g. navigation, relational memory, theory of mind; cf. Cavanna and Trimble 2006). It was also implicated by manipulation of syntactic complexity (Caplan et al., 2003; Shetreet et al., 2007) but, to our knowledge, its role in language processing has never been explicated.

We found the precuneus to be activated by a number of contrast that (we suggest) reflect manipulations of syntactic complexity. Activation in this region positively correlated with the number of parser steps (both in the Word model, 2.8.2 and the
Sentence model, 2.10.2), by the TOM sentential measure (4.7.1) and the belief state parameter (4.7.2) (which most often entails sentential embedding) and by the sentence measure of displacement (5.4.1).

We suggest, then, that the precuneus plays a central role in the representation of hierarchical structure more generally, and linguistic syntactic structure in particular. Data from non-human primate tracing work shows that the precuneus is connected to the precentral gyrus (Cavanna and Trimble, 2006). This connections might support the mapping between linear order and syntactic structure since the precentral gyrus has been suggested to support the representation of temporal order (Vigneau et al., 2006).

Though in the lesion literature there has been a traditional association of syntactic deficits with Broca’s area (BA 44 and BA 45), recent studies have demonstrated that the relevant regions are in fact the adjacent BA 46 and BA 47. Dronkers et al. (2004), in a large scale sentence comprehension study of 64 patients with left hemisphere lesions, found that lesions to BA 46 and BA 47 were associated with impaired comprehension of sentences. Interestingly, only lesions to BA 46 were associated with impaired comprehension of transitive sentences (BA 47 lesions were associated with impaired performance on more complex sentence forms). This region has also been implicated by fMRI studies of sentence comprehension (Ni et al., 2000). In our study, activation in BA 46 correlated with entropy over the beam, or structural ambiguity (3.5.1).

Activation in the posterior middle frontal gyrus (BA 8) correlated positively with the Sentence model syntactic complexity measure (2.10.2). Vigneau et al. 2006 characterize this region (f2p in their notation) as a syntax region, part of the sentence comprehension working memory network. In the motor control literature, It has been associated with planning of complex actions. It is connected to the precuneus via

---

5It is possible, of course, that the same hierarchical representation, namely linguistic syntax, underlies all the different domains mentioned above. We leave this question for further research, however it is interesting to note that both complex navigation and the false belief task have been shown to depend (on-line) on language (Hermer-Vazquez et al., 1999; Newton and de Villiers, 2007).
SLF I and to BA 45 in the inferior frontal gyrus via the LSIF, a connection that is relevant to the role of BA 45 in mediating competition during lexical insertion (or lexical access). In our study, activation in this region correlated with the sentence measure of syntactic complexity.

Having identified the different components of the syntax module it remains to determine the individual role of each region, or node, in that network. Though a precise characterization will need to await future work, we can make some preliminary proposals. BA 46 has been shown to support manipulation of material in working memory (Cabeza and Nyberg, 2000). We propose that BA 46 (probably with the participation of the dorsal striatum; Ullman 2004; Draganski et al. 2008) supports syntactic structure manipulation (structure-building, Merge, and transformations, Re-Merge).

The representation of the syntactic structure itself is supported by the precuneus, which might also serve as an interface between the linguistic syntactic structure and other cognitive faculties or domains such as navigation and episodic memory. Finally, BA 8, which has been implicated by experiments manipulating uncertainty (Volz et al., 2005) might support structural predictions based on the current representation in the precuneus. This interpretation of the role of BA 8 receives further support by the fact that adjacent dorsal prefrontal regions have been found to be activated by lexical surprisal in our experiment.

6.2.3 Spell-out and the Encyclopedia

A number of recent authors have highlighted the role of the ventral frontal - anterior temporal network (mediated via the uncinate fasciculus) in morpho-syntactic processing. Grodzinsky and Friederici (2006) propose that this network is responsible to local phrase structure-building. Marslen-Wilson and Tyler (2007) propose a role for this network in morphological decomposition. Hickok and Poeppel (2007) propose that this network is an interface between the compositional component in the anterior temporal lobe and the articulatory component in the IFG. Vigneau et al. (2006) propose that this network mediates long term memory storage for complex syntactic structure. Our data also demonstrates the syntactic role of this network. Activation
in this network correlated with syntactic surprisal (2.8.2) and entropy (2.10.2) and re-ranking of the alternatives in the beam (3.5.2).

Our proposal is the most similar to that of Vigneau and colleagues. We suggest that this network supports the representation of the output of Spell-out of the syntactic structure. The ventral IFG receives input from the syntactic module via local connections within the IFG. The anterior temporal pole is connected dorsally to auditory (linguistic) regions and ventrally to visual (linguistic) regions (Olson et al., 2007). BA 47 is connected to the articulatory network in the ventral premotor cortex. These connections support the PF interface. The anterior temporal pole is also connected via the ILF to more posterior temporal regions which support semantic representations (Mandonnet et al., 2007).

The same network is also responsible for the long term storage (and retrieval) of a subset of these Spell-out outputs in (and from) long term memory (the Encyclopedia). The uncinate fasciculus also connects the ventral IFG with the hippocampus gyrus which is known to play a role in memory formation and retrieval. This role in storage explains the sensitivity of this system to statistical properties of the syntactic structure (syntactic surprisal).

In on-going work with Roni Katzir on the semantic interpretation at Spell-out, based on earlier work on the topic (Bachrach and Katzir, 2007), we propose that abstraction (binding of a base position of a moved element) takes place at Spell-out. Coupled with the proposal here regarding the network underlying Spell-out, this predicts that the posterior BA 47 will be activated in contrasts that involve syntactic movement. This has been verified by our sentential displacement measure which produced positive activation in this region (??).

6.2.4 The Vocabulary and lexical insertion

There is a general agreement that the the middle temporal gyrus plays a crucial role in lexical processing. In Dronkers et al.'s (2004) large scale study of the effect of left hemisphere lesion on language comprehension, lesion to the MTG (or the white matter underlying it) were associated with word level deficits (not observed for lesions to any
other region). In their meta-analysis, Vigneau and colleagues identify a cluster of semantic related activations in the MTG and describe it as a word processing region. Hickok and Poeppel (2007) make a similar proposal. The meta-analysis of semantic priming by Lau et al. (submitted) has singled out the MTG (for short SOA priming).

Vigneau and colleagues also highlight the location of this region between the ventral visual stream (reading but also other visual information regarding meaning, such as pictures) and the more dorsal auditory stream in the superior temporal gyrus and sulcus. At the same time, much recent work has argued for a distributed view of conceptual knowledge. For example, Hauk et al. (2004) demonstrated that words referring to face, arm, or leg actions (e.g., to lick, pick, or kick), when presented in a passive reading task, differentially activated areas along the motor strip that either were directly adjacent to or overlapped with areas activated by actual movement of the tongue, fingers, or feet.

We have found the MTG to be sensitive to lexical frequency (2.8.1), as well as word length (1.8.2). In chapter 2 we have emphasized the fact that activation in this regions was not sensitive to predictability (lexical surprisal) and interpreted it as evidence for bottom-up lexical access. We suggest that the MTG is the central node in a network of regions which support the Vocabulary, or the stored associations between roots (concepts) and phonological/orthographic form. This proposal is, in essence, a translation of the proposal by Vigneau and colleagues (and in the literature more generally) into the terms of our grammatical architecture. The interface with the phonological form is in the mid superior temporal sulcus (Hickok and Poeppel, 2007). The interface with orthographic form is in the fusiform gyrus (the word form area, Jobard et al. 2003). Conceptual knowledge is stored in distributed fashion throughout the cortex (the inferior and medial temporal lobe, as well as somatosensory regions).

Broca's area (BA 44 and 45) has also been implicated in lexical access or semantic retrieval (Wagner et al., 2001; Lau et al., submitted). This region, and in particular BA 45, is connected to the MTG via the arcuate fasciculus (Catani et al., 2005; Rilling et al., 2008). We propose, following the proposal by Wagner and colleagues, that BA 45 provides top-down modulation of the activation levels of the items in the
Vocabulary. We found activation in this region to correlate with the combined word level model of lexical surprise (2.8.1) and with the sentence level averaged frequency measure (2.10.1). It might not be an essential component of the Vocabulary since lesions to this region do not disrupt word level comprehension. However, it might be critical for the access of the Vocabulary during word production, explaining the deficits often observed in Broca’s aphasia.

Lexical insertion is a related, but distinct, component in our architecture. Lexical insertion is the mapping of particular local syntactic configurations (often limited to a single non-terminal node) onto Vocabulary items. By its nature, lexical insertion is context sensitive since it takes into account grammatical agreement (gender, number, tense, case, etc.) and possibly other contextual information, in the choice of the appropriate Vocabulary item. Vigneau and colleagues identify a semantic cluster in the posterior superior temporal sulcus (PSTS). These authors interpret this regions as an orthography to meaning interface region. We do not dispute this characterization (our data comes from an auditory paradigm so we do not have relevant data); however, we suggest that this region supports (also) lexical insertion. In effect, we found a correlation between activation in this region and lexical surprisal and frequency (2.8.1).

This region is connected to the MTG via local association fibers, to the IFG through the direct route of the arcuate fasciculus, and to the angular gyrus through the indirect route (Catani et al., 2005). In addition, the superior temporal sulcus is connected to the precuneus (Cavanna and Trimble, 2006). Vocabulary items are accessed through the connection with the MTG. The syntactic structure is accessed through the connection with the precuneus. The connection with the IFG provides for contextual top-down modulation of expectations as well as of the competition between Vocabulary items (which is a central component of lexical insertion from the perspective of Distributed Morphology; Halle and Marantz 1993).
6.2.5 Phase integration

One of the central components of the grammatical architecture espoused here is the cyclic application of Spell-out to the syntactic structure. This iterative interpretation of the syntactic structure at each Phase brings about the question of Phase integration. That is, how is the output of a smaller Phase integrated into the output of the Phase containing it? We propose that this integration takes place in the angular gyrus. We have found the angular gyrus to be activated by a number of our measures, including number of derivation steps (2.8.2), TOM (4.7.1) and displacement (5.4.1).

Both imaging (Humphries et al., 2006) and lesion (Dronkers et al., 2004) studies have implicated the angular gyrus in sentence processing. Vigneau et al. (2006) indeed propose that the angular gyrus supports the elaboration of semantic associations (though it is not clear what this process is or how it relates to the manifested role of the angular gyrus in sentence processing). This region has also been implicated by TOM tasks (both by lesion and multiple fMRI studies). Finally, the angular gyrus has been found to be particularly engaged by sentences containing syntactic displacement, both by Dronkers and colleagues' lesion study and fMRI experiments such as Cooke et al. (2002); Constable et al. (2004); Ben-Shachar et al. (2004); Christensen (2008).

The angular gyrus is connected to the MTG and superior temporal regions via the posterior section of the AF (Catani et al., 2005). It is connected to the lateral frontal lobe via the anterior section of the AF (or the SLF III). In addition, it is connected directly to BA 47-anterior temporal network via the inferior longitudinal fasciculus. Non-human primate literature provides evidence for structural connectivity between the inferior lateral parietal (which includes the angular gyrus) and the precuneus. Studies in humans demonstrated strong functional connectivity between the two regions (Marguiles et al., 2008; Hagmann et al., 2008).

The integration of smaller Spell-out objects within larger ones is supported by the connection between the angular gyrus and the BA 47-anterior temporal network as well as the MTG and PSTS. The precuneus connection to the angular gyrus provides
the syntactic frame and the lateral IFG (Broca’s region) provides top-down control.

TOM, as has been emphasized by de Villiers (2007), is intimately associated with
sentential embedding which, in turn, requires Phase integration of the embedded
clause. As expected from the literature, we found angular gyrus activation in the
sentential and word level TOM contrasts.

In non-local dependencies, the reactivation of the displaced constituent in the gap
is, in effect, a case of Phase integration. The only special thing about this integration
is that the Spell-out object has to be retrieved from working memory. This type of
integration arguably requires additional top-down control since the parser must select
the correct target for reactivation. This explains the pattern of activation observed
here for the A-bar gap regressor, which included the angular gyrus (Phase integra-
tion), the hippocampus (retrieval) and Broca’s region (top-down control) (5.4.3).

6.2.6 Propositional representation

A number of experiments which used multi-sentence stimuli (short passages or nar-
ratives) report activation in the anterior BA 47 (Xu et al., 2005; Awad et al., 2007).
In a recent meta-analysis of text comprehension, Ferstl et al. (2008) have found BA
47 to be the the region within the inferior frontal cortex that is activated by tasks
which tap discourse structure. This region has been associated with reward process-
ing; however, Browning and Gaffan (2008) demonstrated that it was not engaged in
all cases of reward processing but only in cases that required a complex (temporal)
representation of the conditions for the reward. In our own work we found activation
within the anterior BA 47 for the sentence level contrasts of syntactic complexity
(2.10.2) and TOM (4.7.1). In addition, the sentence condition itself (whose results
were not reported here) produced activation in that region. We suggest that this
region represents propositional (veri-conditional) meaning (which is associated with
sentences as well as certain tensed complements). Its connections with the limbic
system (e.g. the amygdala) and the medial prefrontal cortex serve as an interface
between language and emotion or action as well as moral reasoning.
6.2.7 Pragmatics: The computation of alternatives

In the heart of the Gricean approach to pragmatic reasoning (Grice, 1975) lies the notion of alternatives. In computing the pragmatic meaning of a sentence, the hearer needs to compare this sentence to a set of other sentences. For example, it is assumed that without additional pragmatic reasoning, the sentence A: ‘John ate some of the cookies’ is compatible with a situation where John ate all of the cookies. However, if one uses this sentence it usually implies that John didn’t eat all of the cookies. This implicature, which is not part of the basic semantic meaning of the sentence (it is not entailed by the meaning of the sentence) is a pragmatic inference.

Grice proposed that this implicature is arrived at via comparison of the original sentence with the alternative sentence B: ‘John ate all of the cookies’. The reasoning (highly simplified) goes as follows: since the speaker did not utter B (which entails A), and since the speaker is assumed to be cooperative and knowledgeable, B must be false (otherwise the speaker would have uttered B instead of A).

Since Grice, the notion of alternatives has become central in the analysis of other aspects of semantic reasoning such as information focus computation (Rooth, 1996) (cf. also Katzir 2008). Moreover, recent work (Spector et al., 2008) suggests that the computation of alternatives is much more local than what was originally proposed and in fact takes place iteratively throughout the semantic computation (possibly at every Phase).

The computation of alternatives requires access, minimally, to the syntactic structure as well as the semantic interpretation of the target linguistic object (Katzir, 2008). In addition, pragmatic computation makes reference to the state of mind or intentions of another and so has a dimension of Theory of Mind. We suggest that BA 9 in the medial prefrontal cortex is engaged by the computation of alternatives. This region has been implicated both by TOM tasks (Spreng et al., 2008) and discourse processing (Ferstl et al., 2008). In effect, Ferstl and colleagues suggest that this overlap might be due to the TOM component of discourse processing, namely pragmatic inferences. This region has also been implicated by moral decision making.
(Heekeren et al., 2003). Pylkkänen and colleagues found the medial prefrontal cortex to be activated in cases of semantic coercion (Pylkkänen, 2008), which might also engage pragmatic reasoning.

This region is connected both the the ventral frontal cortex (BA 47) and the precuneus (Cavanna and Trimble, 2006). These connections supply (respectively) the semantic and syntactic information required for the computation of alternatives. In our own data, the sentence TOM contrast (4.7.1), the sentential combined syntactic measure (entropy and complexity, 2.10.2) as well as the basic sentence condition all overlapped in BA 9. None of these measures is a direct reflection of alternative computation and so further work is required to provide more direct evidence for the proposal here.

6.3 Conclusion

In this chapter we have outlined a neurolinguistic model of grammatical competence. Abstracting away from the specific details, this model highlights a distributed view of the neural implementation of cognitive modules and emphasizes the interface between these different modules. By bringing together functional (and anatomical) data and theoretical linguistic constructs, this model represents an attempt to unify the study of the faculty of language. Needless to say, the proposal here is only a first step towards an eventual full fledged model; however, even at this early stage, such a model can serve to formulate more specific hypothesis regarding the relation between brain activation (or disfunction) and linguistic behavior. We intend to pursue this direction in our future work.
Appendix A

The surprisal and derivation step measures

A.0.1 Notation and preliminaries

Let \( w_1 \ldots w_n \) be an input string of length \( n \). For a given PCFG \( G = (V, T, S^\dagger, P, \rho) \), let \( V \) be the set of non-terminal variables, \( T \) the set of terminal items, \( S^\dagger \in V \) a special start non-terminal, \( P \) a set of rule productions of the form \( A \rightarrow \alpha \) for \( A \in V \), \( \alpha \in (V \cup T)^* \), and \( \rho \) a function that assigns probabilities to each rule in \( P \) such that for any given non-terminal symbol \( X \in V \):

\[
\sum_{\alpha \text{ s.t. } X \rightarrow \alpha \in P} \rho(X \rightarrow \alpha) = 1
\]

For a given rule \( A \rightarrow \alpha \in P \) let the function RHS return the right-hand side of the rule, i.e., \( \text{RHS}(A \rightarrow \alpha) = \alpha \). Without loss of generality, we will assume that for every rule \( A \rightarrow \alpha \in P \), one of two cases holds: either \( \text{RHS}(A \rightarrow \alpha) \in T \) or \( \text{RHS}(A \rightarrow \alpha) \in V^* \). That is, the right-sides include either one terminal item or zero or more non-terminals, but no mixture of non-terminals or terminals on the right-hand side.

We can define a “derives” relation (denoted \( \Rightarrow_G \) for a given PCFG \( G \)) as follows:

---

\(^1\)The majority of this appendix has been written up by Brian Roark
\[ \beta A \gamma \Rightarrow_G \beta \alpha \gamma \] if and only if \( A \rightarrow \alpha \in P \). We say that a string \( w_1 \ldots w_n \) is in the language of a grammar \( G \) if \( S^* \Rightarrow_G w_1 \ldots w_n \). For a given derivation \( S^* \Rightarrow_G \alpha \), where \( \alpha \in (V \cup T)^* \), the series of derivation steps that yield \( \alpha \) can be represented as a tree, with the start symbol \( S^* \) at the root, and the “yield” sequence \( \alpha \) at the leaves of the tree. A complete tree has only terminal items in the yield, i.e., \( \alpha \in T^* \); a partial tree has some non-terminal items in the yield. A leftmost derivation begins with \( S^* \) and each derivation step replaces the leftmost non-terminal \( A \) in the yield with some \( \alpha \) such that \( A \rightarrow \alpha \in P \). Note that with a leftmost derivation, the yield partitions into an initial sequence of terminals \( \alpha_1 \in T^* \) followed by a sequence of non-terminals \( \alpha_2 \in V^* \). For a complete derivation, \( \alpha_2 = \epsilon \); for a partial derivation \( \alpha_2 \in V^+ \).

A leftmost derivation \( D \) of \( m \) steps can be represented as a sequence of rule productions \( d_1 \ldots d_m \), where each \( d_i \in P \) was the rule used at step \( i \) of the derivation. We can define the probability of the derivation \( D = d_1 \ldots d_m \) (and the corresponding tree) as

\[
\rho(D) = \prod_{i=1}^{m} \rho(d_i)
\]  

(A.1)

Let \( D(G, w_1 \ldots w_i) \) be the set of all possible leftmost derivations \( D \) (with respect to \( G \)) such that \( D = d_1 \ldots d_m \) for some \( m \) and \( \text{RHS}(d_m) = w_i \). These are the set of partial leftmost derivations whose last step used a production with \( w_i \) on the right-hand side. Then the prefix probability of \( w_1 \ldots w_i \) with respect to \( G \) is

\[
\text{PrefixProb}_G(w_1 \ldots w_i) = \sum_{D \in D(G, w_1 \ldots w_i)} \rho(D)
\]  

(A.2)

For any derivation \( D = d_1 \ldots d_m \), let \( D^- \) be the derivation without its final step, i.e., \( D^- = d_1 \ldots d_{m-1} \). Then, by definition, \( \rho(D^-) = \rho(D) / \rho(d_m) \).
A.0.2 Surprisal

The surprisal at word $w_i$ is the negative log probability of $w_i$ given the preceding words. Using prefix probabilities, this can be calculated as:

$$S_G(w_i) = -\log \frac{\text{PrefixProb}_G(w_1 \ldots w_i)}{\text{PrefixProb}_G(w_1 \ldots w_{i-1})}$$

(A.3)

Substituting equation A.2 into this, we get

$$S_G(w_i) = -\log \frac{\sum_{D \in \mathcal{D}(G, w_1 \ldots w_i)} \rho(D)}{\sum_{D \in \mathcal{D}(G, w_1 \ldots w_{i-1})} \rho(D)}$$

(A.4)

If we are using a beam-search parser, some of the derivations are pruned away. Let $\mathcal{B}(G, w_1 \ldots w_i) \subseteq \mathcal{D}(G, w_1 \ldots w_i)$ be the set of derivations in the beam. Then the surprisal can be approximated as

$$S_G(w_i) \approx -\log \frac{\sum_{D \in \mathcal{B}(G, w_1 \ldots w_i)} \rho(D)}{\sum_{D \in \mathcal{B}(G, w_1 \ldots w_{i-1})} \rho(D)}$$

(A.5)

Any pruning in the beam search will result in a deficient probability distribution, i.e., a distribution that sums to less than 1. For the particular parser that we are using, Roark’s thesis (2001) showed that the amount of probability mass lost is very low, hence this provides a very tight bound on the actual surprisal given the model.

A.0.3 Lexical and Syntactic surprisal

High surprisal scores result when the prefix probability at word $w_i$ is low relative to the prefix probability at word $w_{i-1}$. Sometimes this is due to the identity of $w_i$, i.e., it is a surprising word given the context. Other times, it may not be the lexical identity of the word so much as the syntactic structure that must be created to integrate the word into the derivations. One would like to tease surprisal apart into “syntactic surprisal” versus “lexical surprisal”, which would capture this intuition of the lexical versus syntactic dimensions to the score. Our solution to this has the beneficial property of producing two scores whose sum equals the original surprisal score.
To calculate the original surprisal score, we created sets of partial derivations at the point when each word $w_i$ is integrated into the syntactic structure, $\mathcal{D}(G, w_1 \ldots w_i)$. We then calculate the ratio from point to point in sequence. To tease apart the lexical and syntactic surprisal, we will consider sets of partial derivations immediately before each word $w_i$ is integrated into the syntactic structure. We already have the apparatus to achieve this set, namely for any $D \in \mathcal{D}(G, w_1 \ldots w_i)$, we consider $D^-$ as defined at the end of Section A.0.1. Using this notation, the syntactic surprisal is calculated as:

$$\text{SynS}(w_i) = -\log \frac{\sum_{D \in \mathcal{D}(G, w_1 \ldots w_i)} \rho(D^-)}{\sum_{D \in \mathcal{D}(G, w_1 \ldots w_{i-1})} \rho(D)} \quad (A.6)$$

and the lexical surprisal is calculated as:

$$\text{LexS}(w_i) = -\log \frac{\sum_{D \in \mathcal{D}(G, w_1 \ldots w_i)} \rho(D)}{\sum_{D \in \mathcal{D}(G, w_1 \ldots w_i)} \rho(D^-)} \quad (A.7)$$

Note that they sum to form total surprisal:

$$\text{SynS}(w_i) + \text{LexS}(w_i) = -\log \frac{\sum_{D \in \mathcal{D}(G, w_1 \ldots w_i)} \rho(D^-)}{\sum_{D \in \mathcal{D}(G, w_1 \ldots w_{i-1})} \rho(D)} - \log \frac{\sum_{D \in \mathcal{D}(G, w_1 \ldots w_i)} \rho(D)}{\sum_{D \in \mathcal{D}(G, w_1 \ldots w_i)} \rho(D^-)}$$

$$= -\log \frac{\sum_{D \in \mathcal{D}(G, w_1 \ldots w_i)} \rho(D^-)}{\sum_{D \in \mathcal{D}(G, w_1 \ldots w_{i-1})} \rho(D)} + \log \frac{\sum_{D \in \mathcal{D}(G, w_1 \ldots w_i)} \rho(D)}{\sum_{D \in \mathcal{D}(G, w_1 \ldots w_i)} \rho(D^-)}$$

$$= -\log \frac{\sum_{D \in \mathcal{D}(G, w_1 \ldots w_i)} \rho(D)}{\sum_{D \in \mathcal{D}(G, w_1 \ldots w_{i-1})} \rho(D)} = S_G(w_i) \quad (A.8)$$

### A.0.4 Derivation steps

Recall that $\mathcal{D}(G, w_1 \ldots w_i)$ is the set of partial leftmost derivations whose last step used a production with $w_i$ on the right-hand side, and $\mathcal{B}(G, w_1 \ldots w_i) \subseteq \mathcal{D}(G, w_1 \ldots w_i)$ represents the set of these derivations still in the beam of the parser. For a given derivation $D$, let $|D|$ be the length of the derivation; i.e., if $D = d_1 \ldots d_m$ then $|D| = m$. For a given set of derivations $\mathcal{D}$, we define the number of derivation steps
$|\mathcal{D}|$ as follows:

$$
|\mathcal{D}| = \sum_{D \in \mathcal{D}} \frac{\rho(D)}{\sum_{D' \in \mathcal{D}} \rho(D')} |D|
$$

For any given derivation $D \in \mathcal{D}(G, w_1 \ldots w_i)$, let $\text{Sub}(D, k)$ be a sub-derivation of $D$ such that $\text{Sub}(D, k) \in \mathcal{D}(G, w_1 \ldots w_k)$ for $k < i$, i.e., if $D = d_1 \ldots d_m$ then $\text{Sub}(D, k) = d_1 \ldots d_n$ for some $n < m$. Then the number of derivation steps required to incorporate word $w_i$ is:

$$
\text{Steps}(w_i) = \sum_{D \in \mathcal{B}(G, w_1 \ldots w_i)} \frac{\rho(D)}{\sum_{D' \in \mathcal{B}(G, w_1 \ldots w_i)} \rho(D')} (|D| - |\text{Sub}(D, i-1)|)
$$
Appendix B

The narratives

B.1 Narrative I

Joe was a big bear of a man, six feet six inches tall and barrel-chested. When he fell off the moon tower, the tremor his fall caused seemed not unlike an earthquake to people who lived by. The stars pierced down that night like tiny pin pricks piercing through black construction paper. Joe looked up at the tower he had finished the day before and wondered who had placed a bucket at the top. It was a bucket he had seen before, he was sure, and he squinted his eyes to try to remember where it had come from. The bucket that Joe was sure he’d seen before was Toni’s bucket. Small and blue, inside its metallic walls it held something unexpected for Joe. Days before, Toni had scaled the moon tower, clutching that very bucket in her right hand. Joe had asked Toni whether she could repair something for him, and the tools that Joe wanted Toni to use were tucked safely into her belt. It wasn’t only Joe’s request for help that had Toni climbing the tower, though: Toni wanted to leave a gift that she knew Joe would like at the very top of the tower. Toni seemed to move slowly at first, but then she gained speed, putting right foot over left, and left over right, again and again. “I want to find something that I can give Joe that will fit in this bucket,” Toni murmured to herself as she climbed up the moon tower. “The only gift that is small enough, and that I think would be good enough for Joe, is a spell.” “This spell will give Joe the thing he wants most in the world,” she thought to herself. The
view that Toni had from the top of the tower was very good. She could see that Joe was out in the field with the horses that he and Toni raised. Suddenly, Toni was shaken from her thoughts by a voice from the fields. “Toni!” yelled Joe. He was leading the horse that Toni loved best. The horse, a gorgeous bay gelding with three white socks, seemed to Toni to be limping, and his head drooped low. Toni managed to scramble down the tower faster than she’d come up. “What did you do to my horse?” she asked, breathless and gasping. “I don’t know,” Joe answered. “Everything was going fine until I rode him over the little hill you found wildflowers on.” The hill, Joe had noticed, was not hard-packed dirt but instead soft sand. It was the sand that he thought had caused the horse’s leg to twist and sprain. “What did I do?” Joe asked, to no one in particular. He kicked the dirt with his left boot, a favorite trick of a discouraged man, but a trick that didn’t do anything to solve the problem. It was unlikely to make the horse any better. The dirt didn’t appear to Toni to cry out. It was probably the sight of her favorite animal turned into a hobbling beast that made Toni, exasperated and heartsick, turn her eyes to the sky and mutter something. Later she wouldn’t be able to recall her exact words, but the meaning was there. The curse had been set, and the spell that Toni had left so carefully for Joe was ruined by that curse. Days passed. The passing of time was not something that either Joe or Toni noticed explicitly, but it was like the dripping of a faucet — unobtrusive, slightly annoying, and indiscernible up to a certain point. The leg that Joe had caused Toni’s beautiful gelding to sprain slowly healed. The horse’s limp became less and less noticeable each day. Toni walked every day from the house to the stable to feed him carrots while scratching his great furry forehead. Joe kept his eyes on the tower. He was waiting for that one night, that one special night on which the stars would align in a near-perfect way. It would be on that night that he would be able to see the constellation that he was looking for. The day on which Joe finished the tower, he beamed at the whole world. Joe watched his tower from the porch all day. It was a beauty — a perfect composite of wood, metal, and concrete. On the night that he had been waiting for, Joe scaled the tower. Something light and flashy caught his eye as he neared the top. The small object was the bucket that Toni had
forgotten about from the day the horse had been hurt. “What in the world... ” he murmured, speaking to no one in particular. The curse Toni had accidentally created took immediate effect. Instead of giving Joe what he most wanted in the world, the spell threw him from his tower. Joe flew through the air, making a near-perfect parabolic arc through the air. If he had been a diver plunging into a swimming pool, the judges would certainly have succumbed and given him a 10. He wasn’t. Legend has it that the tremor from Joe’s fall was felt by neighbors five miles away. Toni to this day believes that the spell she had meant to leave worked, and that Joe wanted to hurt himself by falling off the tower. The fall that Toni concluded Joe had planned to garner sympathy and support for his moon tower became the stuff of stories. Joe broke his pelvis, and cracked three vertebrae in his upper back. Toni to this day wonders why the spell she wanted to please Joe would hurt him so badly, but decides not to think about it too hard. Instead, she chooses to go outside, where the horse that Toni accidentally cursed Joe for waits for their morning ride.

B.2 Narrative II

Julie Andrews’s golden voice once filled this Boston medical clinic. Steven Tyler of Aerosmith has stepped into the exam rooms. And it’s probably the only doctor’s office that holds a grand piano. The silky smooth voice that so many singers appear to desire is the project of the Center for Voice Rehabilitation at Massachusetts General Hospital. What is this place that researchers think will change people’s voices? It’s the treatment the Center provides to celebrities that receives much of the outside attention. But the scientists here are helping all types of people with voice disorders by investigating existing technologies that further research could adapt to vocal problems. The waiting room that patients to the clinic are greeted by is warm and welcoming. On a recent Tuesday afternoon, 12-year-old Gwen Stephens sat with her mother in that waiting room. The book that Gwen seemed to be holding almost like a life raft in her small hands was about Mary Poppins. It was Gwen’s hope that one day she would be able to sing like Julie Andrews. “Many of the outcomes that
we work towards will come when we’re able to develop biological technologies that doctors can use to replace the voice box,” said Dr. Johnson, director of the clinic, when he sat down to speak with Gwen and her mother. The condition doctors have told Gwen she suffers from is caused by nodules, which are painful calluses on her vocal cords. Will the doctors know how to help Gwen? Dr. Johnson showed Gwen how a laser could shave off some of the nodules that make it difficult for her to speak, but the fix would not be permanent. “Currently there is no way to replace any of the tissues that make up the voice box,” Johnson explained. “The substance that I think our labs need to invent would have to do two things well – vibrate like natural tissue, and not break down inside the body. Then, it might work.” But could any substance scientists believe might help people like Gwen really last for a long time inside the body? Researchers do not know exactly how this sort of material would behave. Dr. Johnson’s assistant, Tom, helped Gwen into a chair and told her to lie back. Gwen’s mother watched nervously as Tom seemed to put a tube down Gwen’s throat. “The instruments we’ll be using to see what’s happening inside Gwen’s throat are at the end of this tube,” Tom explained. The machine that Tom hooked the tube up to looked like a large box with a screen on it. “Say ahhhh, Gwen,” said Dr. Johnson. The screen, which was supposed to show the vibrations in Gwen’s throat, blinked several times and then stopped working. Tom was sent away to find the spare batteries that they kept for these occasions. Gwen wondered what was the reason medical equipment always seemed to break when she was around. “Right now, the tissue that we’ve decided to use is from other parts of the body, so it’s different from the tissue that occurs naturally in the throat,” Johnson said. “But some of the materials that the bio-engineers will create will hopefully work like the real thing.” The researchers do not yet know which material the industry needs the scientists to invent, or how to safely make it, but they’re working on many new substances which they hope may prove successful in the future. The implants the researchers are imagining would take the shape of small elastic squares, approximately one half-inch long on either side, and just one millimeter thick, about the thickness of a dime. These implants would be pressed down on the vocal cords like bandages covering the places where
nodules had been removed, and would vibrate normally with the surrounding tissue. Dr. Johnson explained how the implants would work for a patient like Gwen. “The tissues surgeons work with now, we know the limitations of. But who knows what properties the materials we have yet to discover will possess? Anyone who discovers a substance we can use to repair vocal tissue will have the grail. And I think someone will make that discovery in the next five years,” said Johnson, while Gwen and her mother listened attentively. Medical work of the kind that the doctors at the clinic are trying to perform has attracted star power in recent years. In 1997, singer Julie Andrews underwent throat surgery in New York, to remove non-cancerous vocal nodules. The surgery, which is typically minor and relatively safe, went poorly and Andrews lost the ability to sing. The doctors who Andrews alleged to have botched her surgery settled in 2000 for a sum the singer has not disclosed, but that has been estimated at $30 million. Around the same time, Andrews sought out Dr. Johnson for experimental reparative surgery. Now her voice, although not completely healed, seems to her physician to have stabilized. The experimental treatment that was used to repair Andrews voice is heavily regulated by the government. Dr. Johnson can’t speculate about what methods the government will determine that doctors can use in the future, but the options are likely to be greater. For people like Gwen, just having a normal voice would be a gift. “One desire I have is to be able to sing in a choir,” she croaked, after the machine the doctors had used to investigate her nodules was removed from her throat. Gwen’s mother nodded. “If I could, the gift I’d give Gwen would be voice.” Maybe, with the help of the researchers who work at the Voice Rehabilitation Clinic, those dreams will become a reality.

B.3 Narrative III

Harmon was a small and usual-looking boy, who seemed to everyone, including himself, to possess little special talent. He knew this was true because he was often told he was lacking in talent, usually by his master, the Viscount Springiron. Harmon worked in Springirons workshop as an apprentice. The fact that there were other young boys
and even some young girls who the Viscount employed to work there as well, did not seem to matter to the boy. Harmon always felt particularly alone. Perhaps it was the nature of the work—the job that Harmon had performed for a year was screwing small brass winding knobs into the watches Springiron made. Or perhaps it was just that Harmon was a strange little creature. Some of the knobs that Harmon thought had been screwed tightly onto the watches fell off when the watches were passed to the next worker; then the Viscount would scowl and grumble maliciously at Harmon. Shunned by most of the children, Harmon had just one friend in the shop well, near the shop—a pygmy goat named Alexander. He was not the type of goat that ate slop and leaves from trees. No, Alexander seemed to Harmon to prefer the heavy taste of machine grease, which the young boy snuck out to him. The grease that Harmon lovingly slathered on his hands for Alexander to lick was thick and black and smelled slightly sweet. The goat would clean Harmon's hands and then the boy would pat the goats head and belly before returning to work. Harmon sometimes wondered who in the workshop could possible be as good a friend as Alexander the goat was. In the middle of a busy morning of screwing on shiny brass knobs, the Viscount Springiron came to Harmon's desk and cleared his throat. Harmon started to tremble in front of his employer and master. The idea that the master might send him away from the workshop and the goat caused his stomach to do somersaults. The Viscount seemed to wonder how the boy could be so frightened by everything he encountered. “Harmon,” said the Springiron, speaking in a booming voice. “Do you see that boy over there, the one who's shoveling coal?” “Yes,” squeaked Harmon. The boy the master was talking about was blonde and had strong-looking forearms.

“The boy who is shoveling coal right now is going to be helping you with your work from now on instead. Come on, lets meet him.” Harmon obediently followed the Viscount Springiron towards the boy the Viscount had told Harmon he would be working with. The boy had hair the color of sand and a dimple in his left cheek when he smiled, which he was doing as they approached. It was the boys responsibility to be on his absolute best behavior, Harmon knew. “The names Connor,” the boy with the dimpled smile said and held out his hand. The hand that Connor held out was
covered in machine grease. Harmon cringed at the thought of shaking that hand, but he cringed even more when he thought about what the Viscount would do if he didn't shake that hand. The only machine grease that little Harmon thought was okay was the machine grease he loved to feed to Alexander the goat. . “Hi,” Harmon squeaked. “I'm Harmon and I'm a fairly talentless creature.” As soon as the words tumbled out of his mouth, he wondered what he could have done to stop them. But the words that Harmon spoke to Connor were unstoppable. Connor shrugged. “Well, that's alright then. Well find a way to work together.” The instructions that the Viscount gave the boys required that they sit next to each other, so they did. Harmon and Connors responsibility was to screw on the knobs that their master told them would be sitting in the box. They started to work. Other employees who had thought that their own skills were the best in the workshop were amazed by how quickly the two boys worked together from the start and by what number of instruments the two appeared to them to assemble in one day. Each boy had his own responsibility: it was Harmons job to place each brass knob in the proper place on the watch, and then it was Conner who would give the knob a spin to secure it. If the boys didn't form a sturdy friendship, they created something close to one. The work that the Viscount continued to give them became their bond. Soon they were known as the magical boys, Harmon and Connor, or Har-Co, who could make watches faster than anyone else in the workshop. The speed at which the other employees saw them work became legendary. The idea of making teams and partners in the workshop appealed to the Viscount. He originally had only forced Harmon to work with Connor because he didn't approve of the friendship that Harmon had struck up with the goat, but afterwards, as time went on and the rate at which the workshop produced watches increased, he began to wonder which of his other workers he should consider to combine into teams. On a lovely June afternoon, Connor told Harmon that the Viscount said they could take a walk in the sunshine. Later it would turn out that the Viscount had said no such thing, but Harmon didn't care much. The boy with the sand-colored hair and his small dark-featured friend stepped outside into the glittery sunshine and went to play chase with a pot-bellied goat who trotted up to the boys and nudged Harmons
hand. This was the goat named Alexander Harmon had played with before he met Connor. Harmon realized he had a friend wait, two friends.

B.4 Narrative IV

Though the rocks and soil have existed for hundreds of thousands of years, the human story of the island seems likely to have begun in 1386, when the legendary Polynesian adventurer Kokomanu paddled his dugout canoe there. He appeared to be eminently ready for a break, having paddled around the archipelago several times already. He got out of his canoe, stretched, and dropped the garbage he’d been carrying on the ground. As he threw out the peels and skins of fruit he liked to eat to fuel his massive body, Kokomanu accidentally planted the first banana seeds. Before that, the island was just a place for birds to fly over, a punctuation mark of land in the endless stream of blue water. The seeds that the birds which wheeled above watched Kokomanu drop into the clay soil of the island were an omen of things to come. It was those seeds, later on, that seemed to grow with a determination rarely seen outside of humans; even though Kokomanu graced the humid outcropping with his presence for only fifteen minutes, the place was never to be the same again. The island that Kokomanu chanced to discover and then leave was forgotten for just a few more decades. It was several years later when some Dutch traders happened to arrive at the little island. The island seemed to the traders to be uninhabited, and so they thought they were the first ones there. On what did they base this assumption? The traders didn’t see Kokomanu’s careless planting of seeds on the interior of the island. The purpose for which the Dutch traders’ party had come to the island archipelago, of course, was to trade. So, to them, an island devoid of people was not all too useful. The beach they steered towards and landed on was sandy and soft, but traders must trade, so they left their flag and moved on. After the Dutch, explorers from many countries are known to historians to subsequently have moved through the islands. The explorers needed urgently to find food and water after having floated around the enormous blue ocean. Both appeared to the traders to be easily found at the little island, and the
water was drunk and the bananas were eaten. By 1407, it was peasants from the Ly islands to the north who had made their way to settle permanently on the island. The foremost desire of these peasants was to have a place where they could own land, instead of having to work for a master. The houses the peasants thought they built with good materials actually would fall down when earthquakes or hurricanes hit the island, which they did fairly commonly. Some generations passed. The island was sparsely inhabited by the Ly islanders, and there were still boats of European travelers that came by, traded beads for bananas, and left again. The island that Kokomanu preceded the islanders in discovering continued to weather the years gracefully, and was calm and quiet. To protect their houses during storms, the islanders tended to use the leaves of the banana trees. It was this practice that eventually led to a loss of leaves and a smaller annual crop of bananas. Some of the islanders became angry and cursed the god that they believed was controlling the weather. The god that the islanders who were angry cursed was not there, and the trees didn’t seem to the islanders to get any better. The islanders began to fight one another and in 1687, a Portuguese whaling ship floated by. The sight that the whalers who passed by found seemed to confuse them: there was only one banana tree left on the lovely little island, and it was surrounded by the dead or dying islanders. The idea that the Portuguese had was the same as Kokomanu’s: they wanted a quick respite from the ocean. So, they disembarked and decided where to dump their garbage of fruit peels. The island that had already seen so much was about to be cultivated again.
Bibliography


253


cognition and self-referential thought in the medial prefrontal cortex. Journal of
Cognitive Neuroscience 17:1306–1315.

Mason, and M. A. Just. 2008. Predicting human brain activity associated with the

Monsell, S. 1991. The nature and locus of word frequency effects in reading. Basic


Narayanan, S., and D. Jurafsky. 2002. A bayesian model predicts human parse pre-
ference and reading times in sentence processing. Advances in Neural Information

matters for theory of mind, 26–49. Oxford University Press, USA.

based sentence processing classes: Evidence from event-related brain potentials.

Newton, Ashley M., and Jill G. de Villiers. 2007. Thinking while talking: Adults fail

Ni, W., R. T. Constable, W. E. Mencl, K. R. Pugh, R. K. Fulbright, S. E. Shaywitz,
Cognitive Neuroscience 12:120–133.

Ni, W., S. Crain, and D. Shankweiler. 1996. Sidestepping garden-paths: Assessing the
contributions of syntax, semantics and plausibility in resolving ambiguities. lang.

Nicol, J., J. D. Fodor, and D. Swinney. 1994. Using cross-modal lexical decision tasks
to investigate sentence processing. Journal of Experimental Psychology: Learning,
Memory, and Cognition 20:1229–1238.

MSI.

Nobre, A. C., and G. McCarthy. 1995. Language-related field potentials in the
anterior-medial temporal lobe: Ii. effects of word type and semantic priming. Journal
of Neuroscience 15:1090.


272


Zipf, G. K. 1949. *Human behavior and the principle of least effort*. Addison-Wesley Reading MA.