Localising Memory Retrieval and Syntactic Composition: An fMRI Study of Naturalistic Language Comprehension

Shohini Bhattasali* and Murielle Fabre* and Wen-Ming Luh* and Hazem Al Saied* and Mathieu Constant* and Christophe Pallier* and Jonathan R. Brennan* and R. Nathan Spreng and John Hale*

*Co-first authors

Cornell University, Ithaca, NY, USA; Université de Lorraine, CNRS, ATILF, Nancy, France; INSERM-CEA, Paris-Saclay, France; University of Michigan, Ann Arbor, MI, USA; McGill University, Montreal, Canada;

ARTICLE HISTORY
Compiled August 10, 2018

ABSTRACT
This study examines memory retrieval and syntactic composition using fMRI while participants listen to a book, The Little Prince. These two processes are quantified drawing on methods from computational linguistics. Memory retrieval is quantified via multi-word expressions that are likely to be stored as a unit, rather than built-up compositionally. Syntactic composition is quantified via bottom-up parsing that tracks tree-building work needed in composed syntactic phrases. Regression analyses localise these to spatially-distinct brain regions. Composition mainly correlates with bilateral activity in anterior temporal lobe and inferior frontal gyrus. Retrieval of stored expressions drives right-lateralised activation in the precuneus. Less cohesive expressions activate well-known nodes of the language network implicated in composition. These results help to detail the neuroanatomical bases of two widely-accepted cognitive operations in language processing.

KEYWORDS
memory retrieval, structural composition, syntax, lexical cohesion, naturalistic comprehension, fMRI

1. Introduction

Our human ability to comprehend natural language relies on two fundamental of cognitive operations. One involves the retrieval of memorised elements, while the other composes those elements together into novel expressions. Despite a growing body of work on the brain’s language network (see e.g. Dronkers, Wilkins, Van Valin, Redfern, & Jaeger, 2004; Fedorenko et al., 2016; Friederici & Gierhan, 2013; Hagoort & Indefrey, 2013; Pallier, Devauchelle, & Dehaene, 2011) the neural bases of these two operations within this network remain imprecisely specified.

This study disentangles these two cognitive processes with an analysis of fMRI time-courses observed during an extended episode of naturalistic spoken story understanding. Like prior work, it suggests a localisation of memory retrieval and composition to largely distinct regions of the brain. But by linking these operations with quantitatively precise computational models, this study paves the way for more granular research into the relationship between these operations. Further, this spatial dissociation is based upon brain responses to a literary narrative, one that attests a variety of different linguistic phenomena in an ecologically valid setting rather than one or two constructions repeated out of context.

The processing of retrieving memorised elements is operationalised using a specific type of word sequence called “multiword expressions.” This term (henceforth abbreviated MWE)
comes from computational linguistics. Broadly, it refers to word sequences that are better-treated as a single unit rather than as a structurally-composed combination (Sag, Baldwin, Bond, Copestake, & Flickinger, 2002). MWE themselves raise a general question about the balance between productivity and reuse in language processing (Goldberg, 2006; Jackendoff, 2002; O’Donnell, 2015). If they truly lack internal structure, then their processing might proceed via a single, unitary retrieval operation rather than step-by-step composition. Even if they are amenable to some sort of internal structure analysis, existing proposals hold that the entire item is stored and retrieved as a unit.

Adopting this idea as working hypothesis, we assume that MWE comprehension involves an additional memory retrieval, one that is not involved in non-MWE expressions. No assumptions regarding the nature of this retrieval are required, only that the stored element should correspond to the expression as a whole.

This distinction between retrieval and composition figures centrally in neurocognitive models of language processing (Bornkessel-Schlesewsky & Schlesewsky, 2009, 2013; Hagoort, 2005, 2016; Ullman, 2001, 2004, 2015). These models all suppose, to some degree or another, that these two operations in language processing dissociate at the level of neurobiology. The neurobiological implications of this claim are discussed in more detail below in §2.5. The present study adds to prior work by instantiating these two operations in quantitatively precise computational models, and by probing their neural localisation during a naturalistic task.

2. Present Study

We pursue our objective by investigating two specific research questions. The first is to localise these processes (cf. Yang, Marslen-Wilson, & Bozic, 2017): Which brain structures support cognitive processes of structural composition and retrieval of stored linguistic elements, in naturalistic language comprehension? The second research question has to do with the retrieval operation specifically: Are some MWEs better candidates for comprehension via direct retrieval than others?

We investigate these questions using naturalistic spoken story-comprehension, with a view towards to maximising ecological validity. The stimulus, introduced below in §2.1, is a literary text annotated with numbers that quantify processing effort associated with each operation (§2.2 and §2.4). This section finishes by deriving from the existing literature a broad hypothesis regarding the organisation of these two cognitive processes in the brain (§2.5).

2.1. Naturalistic Stimulus

Participants heard de Saint-Exupéry’s The Little Prince (1943) as translated into English by David Wilkinson and read out loud by Nadine Eckert-Boulet. The text constitutes a fairly lengthy exposure to naturalistic language, comprising 15,388 words and lasting over an hour and a half. This follows Willems’ (2015) encouragement to probe the neural bases of language comprehension with greater ecological validity.

The MWE that are identified in this text reflect contemporary usage of these expressions, and they are linguistically diverse (for details see §2.2 and §3.2). To get a sense of how these MWE attestation frequencies relate to distributional patterns in English more generally, Figure 1 compares attestation counts in the stimulus text to their corresponding frequencies in the 560-million word Corpus of Contemporary American English (COCA; Davies, 2008). Over half (56.54%) of the attestations are headed by a verb, and only two items are story-character names. These attestations are evenly distributed throughout the text with about 100 examples in each section.

1 While the experience of listening to an audiobook is quite a natural one e.g. for daily commuters, it is not quite conversational. We put aside the question of which form of language is most basic, contenting ourselves with a comparative sense of the term “naturalistic.” The implicit comparison is between story-listening and lists of unrelated sentence stimuli, devoid of literary content.
2.2. MWEs as Indicators of Memory Retrieval

2.2.1. What are MWEs?

The designation ‘MWE’ is motivated in computational linguistics whenever a sequence of words is better treated as an indivisible whole, rather than a structure that is composed out of sub-parts. This pragmatic criterion (see e.g. Calzolari et al., 2002) picks out a very wide class of linguistic phenomena. To get a sense of this diversity, Table 1 (after Siyanova-Chanturia, 2013) cites some cases that might plausibly be analysed as MWEs. The specific MWE identification procedure used in this study is presented further below in §3.2.

| fixed phrases | per se, by and large |
| noun compounds | black coffee, cable car |
| verb compounds | give a presentation, come along |
| binomials | heaven and hell, safe and sound |
| complex prepositions | in spite of |
| idioms | break the ice, spill the beans |

Table 1. A wide variety of linguistic phenomena are plausibly treated as MWEs.

2.2.2. MWE and Compositionality

What unifies cases of MWE is the absence of a wholly compositional linguistic analysis (Sag et al., 2002). The naturalistic story used as a stimulus in this project attests nearly all of the types cited in Table 1; a few are shown below in Example (1).

(1) a. So I thought a lot about the adventures of the jungle and in turn, I managed with a coloured pencil to make my first drawing.
b. My little fellow, I don’t know how to draw anything except boa constrictors, closed and open.
c. ‘What are you doing there?’ he said to the drinker who he found sitting in silence in front of a number of empty bottles and a number of full bottles.
d. You must see to it that you regularly pull out the baobabs as soon as they can be told apart from the rose bushes to which they look very similar to when they are young.

The syntactic and semantic properties of the boldfaced expressions in (1) indeed resist analysis by parts; they could be considered in some way as lexicalised.

2.2.3. MWE Processing

To the extent that MWEs are not compositional, their processing may depend more on memory retrieval than composition. This idea is most clear in the case of idioms where it has been dubbed the Configuration Hypothesis (CH) by Cacciari and Tabossi (1988).

For the CH, idioms are processed word-by-word, just like any other piece of language, until enough information has accumulated to render the sequence of words identifiable as — or highly expected to be — a memorised chunk. Only at this point the idiomatic meaning is retrieved.

(Cacciari, 2014)

Here, the CH is generalised in supposing that all categories of MWE exhibit this greater dependence on memory retrieval. Section 2.5.1 below, reviews experimental evidence regarding the processing of MWEs. Of course, whether or not retrieval is actually used on a particular occasion may reflect gradations in compositionality. These gradations, as the next section details, can be quantified precisely.

2.3. Lexical Cohesion of MWEs

The crisp theoretical distinction of the CH between compositional and non-compositional expressions obscures the somewhat messier reality: expressions fall along a graded spectrum of compositionality (see Table 2). To quantify this gradient of MWE cohesion, we use a measure called Pointwise Mutual Information (PMI; Church & Hanks, 1990). PMI is commonly used in corpus linguistics to identify collocations.\(^2\) Intuitively, PMI is high when the word sequence under consideration occurs more often together than one would have expected, based on the frequencies of the individual words (Manning & Schütze, 1999 §5.4). More formally, PMI is a log-ratio of observed and expected probabilities:

$$PMI = \log_2 \left( \frac{O}{E} \right)$$

where

$$O = \frac{\text{count(whole expression)}}{\text{corpus size}}$$

and

\(^2\)PMI is one of several cohesion metrics that have been advanced in both computational linguistics and corpus linguistics; Evert (2008) offers a tutorial introduction.
E = \frac{\text{count}(w_1) \cdot \text{count}(w_2) \cdot \ldots \cdot \text{count}(w_n)}{\text{corpus size}^n}

In definition 1 the numerator \( O \) is a probability for the \( n \)-word collocation, \( w_1 w_2 \ldots w_{n-1} w_n \). The denominator \( E \) is what one would expect if the occurrences of each word in the collocation were probabilistically independent.

MWEs that receive a higher PMI score are lexically more cohesive, suggesting less compositionality, at least at the superficial level of word co-occurrence. To the extent that their greater cohesion results in a kind of word-sequence memorisation, these highly-cohesive MWEs are excellent candidates for a processing account based on memory retrieval. Conversely, low PMI scores signal a kind of “incohesion.” These expressions might be better accounted-for by a compositional analysis.

It is important to appreciate that PMI is different from the raw attestation counts that go into it. This is because PMI reflects the ratio of the attestations of the whole relative to the attestations of the parts. It was raw counts, rather than the ratio PMI, that were used in a fMRI study by Yang et al. (2017) that examined memory retrieval in comparison to syntactic complexity. This study found no effects of frequency in the brain beyond the single-word level. Yang et al. (2017) did observe, however, an effect of syntactic complexity by parametrically varying six types of nominal, verbal and prepositional phrases. The next section presents an alternative way to investigate this same factor using naturalistic text.

2.4. Bottom-up Parsing as Composition

To study composition itself, some independent characterisation of compositional processes themselves are needed. In this study, bottom-up parsing plays that role. The intermediate states of this parsing strategy quantify the amount of compositional work that an idealised system would do, in the course of processing the naturalistic stimulus text introduced above in \( \S \) 2.1. While there exists a large literature on bottom-up parsing within computational linguistics (see e.g. Hale, 2014) its essential character is easy to grasp.

Bottom-up parsing amounts to a repeated cycle of choice: whether to shift to the next word or reduce a sequence of transient elements held in memory. As shown in Figure 2, reduce actions are individuated by particular grammar rules. The number of parser actions required at each word defines an incremental complexity metric. We use this complexity metric to quantify

<table>
<thead>
<tr>
<th>PMI</th>
<th>multiword expression receiving this score</th>
</tr>
</thead>
<tbody>
<tr>
<td>26.59</td>
<td>heart skipped a beat</td>
</tr>
<tr>
<td>23.80</td>
<td>have nothing to do with</td>
</tr>
<tr>
<td>21.26</td>
<td>forehead with a handkerchief</td>
</tr>
<tr>
<td>21.18</td>
<td>burst into tear</td>
</tr>
<tr>
<td>20.17</td>
<td>once upon a time</td>
</tr>
<tr>
<td>20.15</td>
<td>boa constrictor</td>
</tr>
<tr>
<td>18.85</td>
<td>peal of laughter</td>
</tr>
<tr>
<td>-2.34</td>
<td>be order</td>
</tr>
<tr>
<td>-2.49</td>
<td>do calculation</td>
</tr>
<tr>
<td>-2.72</td>
<td>be object</td>
</tr>
<tr>
<td>-2.98</td>
<td>be hundred</td>
</tr>
<tr>
<td>-3.15</td>
<td>a well</td>
</tr>
<tr>
<td>-3.50</td>
<td>drink anything</td>
</tr>
<tr>
<td>-3.63</td>
<td>have plan</td>
</tr>
</tbody>
</table>

Table 2. Graded spectrum of compositionality revealed by pointwise-mutual information.
composition effort in the brain, following prior work in neurolinguistics (Brennan et al., 2012; Brennan & Pylkkänen, 2017; Brennan, Stabler, Van Wagenen, Lühi, & Hale, 2016; Nelson et al., 2017) and psycholinguistics (Frazier, 1985).

Figure 2. Panel (A) depicts hierarchical structure for *John loves Mary* to be recognised via processes of syntactic composition with the word-by-word parser action counts given in orange. Panel (B) shows the sequences of parser actions (i.e. **shift** and **reduce**) that would build the colour-coded tree nodes during bottom-up parsing.

### 2.5. Hypotheses regarding Retrieval of MWEs and Composition

With operational definitions sketched-out for both memory retrieval (§2.2) and composition (§2.4), this section turns to what is already known about their processing in the brain. Synthesising several different literatures, a general prediction emerges that these two processes are subserved by different brain areas.

#### 2.5.1. MWE Processing is Different: Experimental Evidence

Prior work supports the claim that MWE comprehension is distinct from other language processing such as structural composition and involves additional memory retrieval. For instance, it is well-established at the behavioural level that MWEs are produced and understood faster due to their frequency, familiarity, and predictability (Siyanova-Chanturia & Martinez, 2014). This would follow if MWEs were remembered as chunks, in the sense of Miller (1956) that was later formalised by Laird, Rosenbloom and Newell (1986, 1987).

Eye-tracking and EEG work further documents this processing advantage across a wide range of MWE sub-types, e.g. binomials (Siyanova-Chanturia, Conklin, & Schmitt, 2011), phrasal verbs (Yaneva, Taslimipoor, Rohanian, & An Ha, 2017), complex prepositions (Molinaro, Canal, Vespignani, Pesciarelli, & Cacciari, 2013; Molinaro, Vespignani, Canal, Ponsa, & Cacciari, 2008), nominal compounds (Molinaro & Carreiras, 2010; Molinaro, Carreiras, & Duñabeitia, 2012), lexical bundles (Tremblay & Baayen, 2010; Tremblay, Derwing, Libben, &
Westbury (2011), and idioms (Laurent, Denhières, Passerieux, Iakimova, & Hardy-Baylé, 2006; Rommers, Dijkstra, & Bastiaansen, 2013; Siyanova-Chanturia, Conklin, & Van Heuven, 2011; Strandburg et al., 1993; Underwood, Schmitt, & Galpin, 2004; Vespignani, Canal, Molinaro, Fonda, & Cacciari, 2010).

2.5.2. Neural Basis for Retrieving MWEs and Composition

The hypothesis that MWEs involve additional memory retrieval reflects trends in both Construction Grammar and Connectionism (see e.g. Ellis, 2008). In these quarters, however, the relevant ‘memory’ does not receive a precise psychological or neural characterisation. It may be that memory for MWE fall into a distinct theoretical category “between” episodic memory and general semantic memory (Renoult, Davidson, Palombo, Moscovitch, & Levine, 2012). Memories in this category are encoded on the basis of repeated personal experience, an idea that is certainly plausible for MWEs. Neuroanatomical sites for the category of memory that Renoult et al. (2012) discuss include medial temporal regions and the Precuneus. This last suggestion converges well with single word retrieval (Halsband, Krause, Sipilä, Teräs, & Laihinen, 2002).

Regarding composition operations, one prominent line of work implicates frontal regions, such as the pars triangularis and pars opercularis of the inferior frontal gyrus (Friederici & Gierhan, 2013; Hagoort, 2016; Sijnders et al., 2009; Zaccarella & Friederici, 2017). Other lines of investigation implicate anterior temporal regions, based on deficit-lesion data (Dronkers et al., 2004) and text comprehension tasks (Ferstl, Neumann, Bogler, & Von Cramon, 2008). These anterior temporal areas are sensitive to parametric variation of phrase size, which would follow if they were somehow representing composed structures (Bemis & Pylkkänen, 2011a; Brennan et al., 2012; Pallier et al., 2011; Pylkkänen, 2016). These latter findings, although, conflict with patient studies showing that atrophy to anterior temporal regions does not systematically impair processing of syntactically complex sentences (Wilson et al., 2014a). This tension between functional and deficit/lesion results remains to be resolved.

This array of empirical work, in combination with the neurocognitive models of language processing discussed below (2.5.3) serves to motivate the present investigation into the brain bases for MWE processing and compositional processing.

2.5.3. Neurocognitive Models of Language and Localisation

Current neurobiological models for language suggest that retrieval and composition should be sub-served by different brain areas. These models do however differ both in their conceptualisation of these operations, and on their anatomical localisation.

The Declarative-Procedural model (Ullman, 2001) for instance, is founded upon a distinction that contrasts memory-related with non-memory-related processing. Ullman (2015) links rule-based mechanisms to frontal regions and sub-cortical structures, while memory for words is supported by medial temporal regions.

In Hagoort’s Memory, Unification and Control framework (2016), composition falls under the scope of the Unification operation and is assigned to inferior frontal areas. While operating under different assumptions, this localisation is in general agreement with Zaccarella and Friederici (2017) who propose that hierarchical processing is subserved by a sub-part of the left inferior frontal gyrus. Regarding the Memory aspect of their model, Hagoort and colleagues agree partly with Ullman, associating that function (among others) to posterior temporal regions (Hagoort, 2009; Hagoort & Indefrey, 2014).

The Dual Streams Model (Hickok & Poeppel, 2007) similarly locates the Lexical Interface, where individual words would be processed, to posterior middle temporal gyrus. Syntactic phrases would be composed, part-by-part, by a combinatorial network within the anterior temporal lobes.

Another perspective is offered by the Extended Argument Dependency Model (eADM; Bornkessel-Schlesewsky & Schlesewsky, 2009, 2013), that divides up language processing in
a different way. In this model sequential information (for instance about word order) is handled by a dorsal stream, while dependency information (as expressed through case-marking) is handled by a ventral stream. If MWE comprehension is sequential processing in this sense, then structures along this dorsal stream, including the inferior parietal lobule, should be involved. On the other hand, composition should activate temporal regions along the ventral stream.

These models’ localisation claims contrast in detail with each other, and with an important body of evidence from patient work (e.g., Dronkers et al. 2004; Wilson et al. 2014b; Wilson, Galantucci, Tartaglia, & Gorno-Tempini 2012, inter alia). Despite this tension, they agree on the common proposal that retrieval and composition should each be independently localisable. It is this point of agreement that motivates the present study.

3. Methods

A spoken narrative serves as the stimulus. Participants hear this narrative over headphones while they are in the scanner (e.g. Brennan et al. 2012; Willems, Frank, Nijhof, Hagoort, & van den Bosch 2015). The sequence of neuroimages collected during their session becomes the dependent variable in a regression against word-by-word predictors, derived from the text of the story. The overall approach to deriving time series predictions regarding the comprehension of this auditory stimulus is shown in Figure 3.

3.1. Subjects and Design

Participants were fifty-one volunteers (32 women and 19 men, 18-37 years old) with no history of psychiatric, neurological, or other medical illness or history of drug or alcohol abuse that might compromise cognitive functions. All strictly qualified as right-handed on the Edinburgh handedness inventory Oldfield (1971). They self-identified as native English speakers and gave their written informed consent prior to participation, in accordance with Cornell University IRB guidelines.

3.2. Stimulus and MWE Identification

The audio stimulus was a literary text, Antoine de Saint-Exupéry’s The Little Prince. We applied two different approaches to MWE identification, which together yield 669 distinct MWE types for a total of 1292 attestations in the stimulus text. This section explains the two identification approaches.

The first approach uses ltagger (Constant & Sigogne, 2011), a freely-available program that combines two submethods:

1. string-based matching to look up MWE in external dictionaries
2. a tagging model, based on conditional random fields and trained on hand-checked examples

Regarding submethod 1, the external dictionaries are: the Unitex lexicon (Paumier, Nakamura, & Voyatzis 2009), the SAID corpus (Kuiper, McCann, Quinn, Aitchison, & van der Veer 2003), the Cambridge International Dictionary of Idioms (White 1998), and the Dictionary of American Idioms (Makkai, Boatner, & Gates 1995).

Regarding submethod 2, conditional random fields are used to guess MWE marks like those shown in the final column of Table 3 (N. Smith 2011, §3.5) offers a general introduction to the technique. The key idea here is to construe a word’s MWE status as a latent variable that is related to the surface string via a collection of weak clues, the “features.” Here the training examples come from MWE annotations in the English Universal Dependencies English Web Treebank (Silveira et al. 2015). This annotated data comprises several web-related genres including weblogs, newsgroups, emails, reviews, and Yahoo! answers. The conditional random field makes its marking decisions based on feature templates that specify, for example, particular lexical
Figure 3. Deriving an expected blood-oxygen level dependent (BOLD) signal from a naturalistic text. (A) shows a segment of the spoken stimulus, with word boundaries in light blue. (B) highlights multiword expressions (MWEs), which may be retrieved directly from memory during comprehension and (C) depicts the phrase structure of the example sentence, which is composed part-by-part during comprehension. In (D) the value of these word-by-word predictors are shown together. The orange values reflect steps taken between each word by a bottom-up (BU) parser. The purple values reflect PMI scores, which quantify the cohesion of MWEs like *has to* and *look after*. Panel (E) shows the expected BOLD response, after these predictors are convolved with a haemodynamic response function (HRF).

<table>
<thead>
<tr>
<th>Idx</th>
<th>Word form</th>
<th>Lemma</th>
<th>POS</th>
<th>MWE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>I</td>
<td>i</td>
<td>PRP</td>
<td>-</td>
</tr>
<tr>
<td>2.</td>
<td>was</td>
<td>be</td>
<td>VBD</td>
<td>-</td>
</tr>
<tr>
<td>3.</td>
<td>thinking</td>
<td>think</td>
<td>VBG</td>
<td>-</td>
</tr>
<tr>
<td>4.</td>
<td>of</td>
<td>of</td>
<td>IN</td>
<td>-</td>
</tr>
<tr>
<td>5.</td>
<td>the</td>
<td>the</td>
<td>DT</td>
<td>+</td>
</tr>
<tr>
<td>6.</td>
<td>other</td>
<td>other</td>
<td>JJ</td>
<td>+</td>
</tr>
<tr>
<td>7.</td>
<td>day</td>
<td>day</td>
<td>NN</td>
<td>+</td>
</tr>
<tr>
<td>8.</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Training data for MWE identification systems. Each word of the sentence is on a separate line with its lemma and its POS tag. + and − indicate whether the word is part of a MWE or not.

items, capitalisation, or part of speech tags. A complete table of these templates is presented in Constant and Tellier (2012). The approach is limited to contiguous MWEs and identifies mainly fixed expressions such as complex function words and nominal MWEs. Crucially, this way of
identifying MWEs is blind to hierarchical syntactic structure.

The second approach uses a transition based system (Al Saied, Candito, & Constant 2017). This system is a variant of the well-known Nivre parser (Constant & Nivre 2016), in which abstract “actions” update an abstract computational state that moves through the text, emitting MWE marks as a side-effect. The choice of which action to take is made by a classifier, here a support vector machine (for textbook introductions see e.g. Abney 2007, §6.4 or Murphy 2012, §14.5). There are actions that Add or Remove words from the computational state, as well as actions which Mark candidates as MWEs. This last type is actually divided into two cases, one of which composes candidate MWEs in a binary fashion, and another which marks these binary elements in a way that produces output. Figure 4 shows how this transition-based system would apply to the example “see to it.”

The input to this classifier includes word form, lemma and part-of-speech information as shown in Table 3, as well as dictionary-based features and features that relate to previous actions (“history-based”). The classifier’s output is a selection of one particular action. This second way of identifying MWEs is similarly denied access to any hierarchical syntactic information.

The training data for this transition-based system come from the Children’s Book Test (Hill, Bordes, Chopra, & Weston 2015), which is close to the genre of the stimulus narrative. Training examples were generated by string-matching based on the external dictionaries cited above. Compared to the conditional random field-based approach, this second approach seems betterable to find verbal MWE.

The MWEs used in the fMRI study reported here were identified by applying both of the approaches described above to the stimulus text. The union of their outputs yielded a time series of indicator variables as shown in Panel C of Figure 5. This time series has the value 1 on words that are final in a MWE, and 0 otherwise.

PMI scores for each MWE define a gradient metric and were determined using equation 1. The probabilities used in this calculation were estimated using the Corpus of Contemporary English (Coca; Davies 2008). The observed probability $O$ of entire MWEs was estimated by querying the 2018 version of COCA, which contains 560 million words. The expected probabilities $E$ are based on an earlier COCA release that comprises 450 million words. These estimates are based on counts of lemmas, rather than inflected words. This use of lemmas serves to focus the analysis on the cohesion of the expression rather than its morphology.

3.3. Annotating compositional structure

Compositional processing was modeled as bottom-up parsing, as introduced above in §2.4. The tree structures were obtained using Stanford parser (D. Klein & Manning 2003) and follow the conventions of the Penn Treebank (Marcus, Marcinkiewicz, & Santorini 1993). They capture constituency relationships, including phrase type and embedding, but do not explicitly mark long-distance dependencies. The complexity metric is defined as the number of reduce actions that would be taken during bottom-up parsing of these trees. Since narrative stimulus does not readily lead to mis-understandings or “garden-path” effects in the sense of Bever (1970) we restrict ourselves to just the globally-correct parser action sequence.

Figure 5 illustrates the two retrieval estimators and the one compositional estimator for a single sentence from The Little Prince. Correlation matrices for all estimators are given in the Appendix.

3.4. Data Presentation

After giving their informed consent, participants were familiarised with the MRI facility and assumed a supine position on the scanner gurney. The presentation script was written in PsychoPy (Peirce 2007). Auditory stimuli were delivered through MRI-safe, high-fidelity head-phones (Conlon HP-VS01, MR Confon, Magdeburg, Germany) inside the head coil. The headphones were secured against the plastic frame of the coil using foam blocks. Using a spoken
Figure 4. A transition-based approach to finding the multiword expression see to it from example (1-d). Identification involves adding see, to and it to consecutive computational states, creating a binary element and marking this element as a multiword expression. For each state, the classifier must choose the appropriate action among the four available actions. Panel (A) shows the effect of some features on choosing the action Add instead of the action Remove for the word “to”. Panel (B) shows the effect of choosing Mark instead of Remove for the composed element “see to it”. Panel (C) shows the effect of some other features on choosing the action Remove instead of Add for the word “that”. These feature templates help the classifier choose the right action. For display purposes we show feature weights as they would be used in a simpler Logistic Classifier, rather than the actual support vector-machine (SVM) that is used in the model. In this simplified setup, decisions are made by passing the summed feature weights – the coloured bars – through a logistic function. The SVM decision rule is more complicated. For details, see Al Saied et al. (2017).
recitation of the US Constitution, an experimenter increased the volume until participants reported that they could hear clearly. Participants then listened passively to the audio storybook for 1 hour 38 minutes. The story was divided into nine sections and at the end of each section the participants were presented with a multiple-choice questionnaire with four questions (36 questions in total), concerning events and situations described in the story. These questions served to confirm participants’ comprehension. They were viewed via a mirror attached to the head coil and answered through a button box with their right hand. The entire session lasted around 2.5 hours.

3.5. Data Acquisition

Imaging was performed using a 3T MRI scanner (Discovery MR750, GE Healthcare, Milwaukee, WI) with a 32-channel head coil at the Cornell MRI Facility. Blood Oxygen Level Dependent (BOLD) signals were collected using a T2-weighted echo planar imaging (EPI) sequence (repetition time: 2000 ms, echo time: 27 ms, flip angle: 77deg, image acceleration: 2X, field of view: 216 x 216 mm, matrix size 72 x 72, and 44 oblique slices, yielding 3 mm isotropic voxels). Anatomical images were collected with a high resolution T1-weighted (1 x 1 x 1 mm³ voxel) with a Magnetisation-Prepared RApid Gradient-Echo (MP-RAGE) pulse sequence.

4. Data Analyses

4.1. Preprocessing

FSL’s Brain Extraction Tool [Jenkinson, Beckmann, Behrens, Woolrich, & Smith, 2012, S. M. Smith et al. 2004] was used for skull-stripping with a fractional intensity threshold.
setting of 0.5. Subsequent preprocessing steps were carried out using AFNI version 16 (Cox, 1996). Anatomical and functional images were co-registered using the in-built AFNI function 3dseg, images were normalised to the MNI–152 template, and images were resampled to 2mm isotropic voxels.

We used multi-echo independent components analysis (ME-ICA) (Kundu et al., 2013; Kundu, Inati, Evans, Luh, & Bandettini, 2012) to improve the signal-to-noise ratio in these data. ME-ICA splits the $T2^*$ signal into BOLD-like and non BOLD-like components. Removing these non-BOLD components mitigates noise due to participants’ head motion, physiology and scanner conditions such as thermal changes (Kundu et al., 2017). Indeed, there were no exclusions based on degree of head movement. Nor was any high-pass filtering or smoothing applied at this stage.

4.2. Statistical Analysis

The research questions layed out above in §2 motivate two statistical analyses. The first analysis localises composition and memory-retrieval operations during naturalistic listening. The second analysis investigates multiword expressions along a quantitative gradient of cohesion. Both analyses employ the General Linear Model, and were carried out using SPM12 (Friston, Ashburner, Kiebel, Nichols, & Penny, 2007). The predictors were convolved using SPM’s canonical HRF.

4.2.1. Analysis 1: With Categorical Predictors

We regressed the word-by-word predictors described below against fMRI timecourses recorded during passive story-listening in a whole-brain analysis. Along with the parser action count and MWE indicators of theoretical interest, four “nuisance” variables of non-interest were entered into the GLM analysis, including a unigram word frequency predictor based on attestations in movie subtitles (Brysbaert & New, 2009). This enables the effects of MWE, Parser-action count, and MWE cohesion to be assessed in way that is distinct from single-word frequency effects. We also included two variables that account for prosodic effects: the RMS amplitude and the fundamental frequency of the auditory stimulus. These control variables serve to improve the sensitivity, specificity and validity of activation maps (Bullmore et al., 1999; Lund, Madsen, Sidaros, Luo, & Nichols, 2006). The predictors entered into Analysis 1 are summarised in Table 4. Figure 5 shows the theoretical predictors’ time series for a particular stimulus sentence.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bottom-up parser action count</td>
<td>Number of REDUCE actions taken since the last word (§2.4)</td>
</tr>
<tr>
<td>Categorical MWE predictor</td>
<td>1 at the last word of a MWE, 0 otherwise (§2.2)</td>
</tr>
<tr>
<td>Word rate</td>
<td>Indicator for spoken word offset</td>
</tr>
<tr>
<td>Word frequency</td>
<td>log-frequency in movie subtitles (Brysbaert &amp; New, 2009)</td>
</tr>
<tr>
<td>f0</td>
<td>fundamental frequency of the narrator’s voice, reflects pitch intensity, an acoustic correlate of volume</td>
</tr>
</tbody>
</table>

Table 4. Predictors used in Analysis 1

4.2.2. Analysis 2: With Gradient Predictors

Analysis 2 uses the same predictors as in Analysis 1, except that the categorical indicator of MWE presence is replaced with a gradient predictor. As explained above in §2.3, this predictor rates the propensity of a MWE to be memorised or not using a standard measure of collocation strength called pointwise mutual information (PMI). The same MWEs that were annotated with a value of “1” in Analysis 1 are in Analysis 2 marked with their PMI score. This is shown in the bottom row of Figure 5.

Note that parser action count is not correlated with the categorical nor with the gradient
MWE predictor \( (r = 0.09 \text{ in both cases}) \). Correlation matrices for all terms entered into the regression analyses are given in the Appendix.

### 4.2.3. Group-level Analysis

In the second-level group analysis, each contrast was analysed separately at the group-level. An 8 mm FWHM Gaussian smoothing kernel was applied on the contrast images from the first-level analysis to counteract inter-subject anatomical variation. All the group-level results reported in the next section underwent FWE voxel correction for multiple comparisons which resulted in T-scores > 5.3.

### 5. Results

Behavioural results of the comprehension task showed attentive listening to the auditory story presentation. Across 51 participants, average accuracy on the comprehension questions was 90\% (SD = 3.7\%). All whole-brain effects reported survived a \( p < 0.05 \) Family-Wise-Error threshold at the voxel level. Tables introduced below use brain region labels from the Harvard-Oxford Cortical Structure Atlas.

#### 5.1. Analysis 1: Results with Categorical Predictors

##### 5.1.1. Results for Composition

Bottom-up parser action count shows a broad activation pattern both in right and left hemisphere. The peak activation is right lateralised in the anterior temporal lobe within a main cluster of activation which extends through the middle and superior temporal gyri. While anterior temporal activation is bilateral, both middle temporal gyrus and posterior superior temporal gyrus are only right lateralised. The second strongest cluster of increased activation is observed in the left inferior frontal gyrus stretching over pars orbitalis and triangularis and extending to the anterior insula and the putamen. A similar increased activation is observed in the right inferior frontal gyrus.

Additional activation clusters were seen in areas including the prefrontal cortex, supplementary motor areas, the cerebellum, the left temporal-parietal junction, as well as inferior temporal lobes. These are all detailed in Table 5 and Figure 6.
Table 5. Significant clusters of increasing activation for bottom-up parser action count after FWE voxel correction for multiple comparisons with p < 0.05 and cluster-extent threshold (k > 50). Peak activation is given in MNI Coordinates, and brain region labels come from the Harvard-Oxford Cortical Structure Atlas.

Table 6. Significant clusters of increasing activation for multiword expressions after FWE voxel correction for multiple comparisons with p < 0.05. Peak activation is given in MNI Coordinates.

5.1.2. Results for MWE Presence

The categorical MWE predictor gives rise to two clusters of activation both in the right precuneus cortex, as presented below in Figure 6 and Table 6.

5.2. Analysis 2: Results with Gradient Predictors

Analysis 2 investigates memory retrieval further by rating how cohesive each MWE is. This is done using pointwise mutual information, introduced above in §2.3 Increasing MWE cohesion, as seen through the positive correlation with PMI, yields a single cluster in the right precuneus.

Left-lateralised activity in superior frontal gyrus, angular gyrus, pars triangularis, posterior middle temporal gyrus, and frontal pole was observed in proportion to decreasing lexical cohesion, as seen through the negative correlation with PMI scores. These are detailed in Table 7 and in Figure 7.
Figure 6. Whole brain contrast images with significant clusters are projected onto a template brain (Holmes et al. 1998). Panel (A) shows us the significant clusters Bottom-up parser action count in orange; Panel (B) shows the significant clusters for Multiword expressions in blue. All images are underwent FWE voxel correction for multiple comparisons with $p < 0.05$. A cluster-extent threshold $k > 50$ is applied for display purposes.
<table>
<thead>
<tr>
<th>Regions for PMI</th>
<th>Cluster size (in voxels)</th>
<th>MNI Coordinates x y z (corrected)</th>
<th>p-value</th>
<th>T-score (peak level)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Correlated with increasing MWE cohesion</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R Precuneus</td>
<td>244</td>
<td>6 -68 56</td>
<td>0.000</td>
<td>7.33</td>
</tr>
<tr>
<td><strong>Correlated with decreasing MWE cohesion</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L Superior Frontal Gyrus</td>
<td>2039</td>
<td>-18 -32 52</td>
<td>0.000</td>
<td>8.39</td>
</tr>
<tr>
<td>L Precentral Gyrus (BA9)</td>
<td>-44</td>
<td>8 40</td>
<td>0.000</td>
<td>7.26</td>
</tr>
<tr>
<td>L Middle Frontal Gyrus</td>
<td>-38</td>
<td>22 46</td>
<td>0.000</td>
<td>6.89</td>
</tr>
<tr>
<td>L Angular Gyrus</td>
<td>688</td>
<td>-42 -58 34</td>
<td>0.000</td>
<td>7.27</td>
</tr>
<tr>
<td>L Inferior Parietal Lobule</td>
<td>-48</td>
<td>-46 50</td>
<td>0.000</td>
<td>5.76</td>
</tr>
<tr>
<td>L Inferior Temporal Gyrus</td>
<td>320</td>
<td>-60 -44 -4</td>
<td>0.000</td>
<td>7.25</td>
</tr>
<tr>
<td>L Inferior Frontal Gyrus Pars Triangularis</td>
<td>211</td>
<td>-46 30 18</td>
<td>0.000</td>
<td>6.41</td>
</tr>
<tr>
<td>L Middle Temporal Gyrus</td>
<td>152</td>
<td>-56 0 -32</td>
<td>0.001</td>
<td>6.49</td>
</tr>
<tr>
<td>L Frontal Pole (BA10)</td>
<td>50</td>
<td>-6 64 20</td>
<td>0.001</td>
<td>5.88</td>
</tr>
<tr>
<td>R Superior Frontal Gyrus</td>
<td>35</td>
<td>14 52 28</td>
<td>0.001</td>
<td>6.15</td>
</tr>
<tr>
<td>L Inferior Frontal Gyrus orbitalis</td>
<td>33</td>
<td>-38 48 18</td>
<td>0.001</td>
<td>5.55</td>
</tr>
<tr>
<td>R Inferior Temporal Gyrus</td>
<td>21</td>
<td>58 -10 -32</td>
<td>0.001</td>
<td>5.78</td>
</tr>
<tr>
<td>R Superior Frontal Gyrus/ SMA (BA6)</td>
<td>15</td>
<td>12 24 58</td>
<td>0.001</td>
<td>5.94</td>
</tr>
</tbody>
</table>

Table 7. Significant clusters for the increasing and decreasing cohesion measure of MWEs after FWE voxel correction for multiple comparisons with \( p < 0.05 \). Peak activation is given in MNI Coordinates.

Figure 7. Significant clusters for the increasing and decreasing cohesion measure of MWEs after FWE voxel correction for multiple comparisons with \( p < 0.05 \). Increasing cohesion measures are represented in blue and decreasing cohesion measures are represented in orange.

6. Discussion

This study investigated the neural substrates of two processes in language comprehension: retrieval and composition of stored linguistic elements. Departing from prior work, we distinguish these operations using computational methods and a naturalistic stimulus. Multiword expressions identified in The Little Prince using the methods of Al Saied et al. (2017) and Constant and Sigogneau (2011) serve as a hypothesis about points in the narrative where extra memory retrievals would occur.

6.1. Lexical Retrieval and MWEs

Significant activation in the precuneus for MWEs in both Analysis 1 and Analysis 2 offers initial support for the idea that this region subserves memory retrieval of these stored elements.
6.1.1. Precuneus and Retrieval of MWEs

Both the categorical Analysis 1 and the gradient Analysis 2 indicate the centrality of the precuneus for MWE-related memory processes. This area features a graded effect such that it is more active for increasingly cohesive MWEs, as estimated by the PMI metric. As discussed above in §2.3 highly-cohesive MWEs are excellent candidates for a processing account based on direct retrieval from memory.

The involvement of the precuneus in the sort of memory retrieval prompted by MWEs is consistent with studies attesting its participation in larger memory networks, e.g., for verbal material (Halsband et al., 2002). The functional characterisation of the precuneus as part of a network sub-serving memory tasks, has been reported for different memory-based processes, such as verbal memory (Halsband et al., 2002), spatial memory (Wallentin, Weed, Østergaard, Mouridsen, & Roepstorff, 2008), episodic memory (Andreasen et al., 1995), memory-related imagery (Fletcher et al., 1995; Mashal, Vishne, & Laor, 2014). Notably, some studies found that the Precuneus is active in non-imagery related episodic memory for musical sequences or abstract words (e.g. Platel et al., 1997).

The precuneus has also been identified as a part of the Default Network, perhaps playing a role in a Dorsal-Medial subsystem of that network (Andrews-Hanna, Smallwood, & Spreng, 2014). This subsystem is said to support story comprehension among other aspects of self-generated thought. Although the precuneus has been designated as part of the Protagonists Perspective Interpreter Network (Mason & Just, 2006) and attested in naturalistic reading by Wehbe et al. (2014), an interpretation of the results of this particular study in terms of reference to story characters is implausible. This is because less than 2% of the MWEs in the stimulus narrative are names of story characters.

More broadly, the postero-medial portion of the parietal lobe has been widely linked to processing of complex lexical information by previous studies. Its sensitivity to the number of complements of a verb was reported by Shetreet, Palti, Friedmann, and Hadar (2007) in the medial precuneus and the anterior cingulate cortex (see also den Ouden, Fix, Parrish, & Thompson, 2009; Shetreet, Friedmann, & Hadar, 2010). The precuneus has also been linked to lexical processing of information of a relatively high complexity (Shetreet, Friedmann, & Hadar, 2009). These results converge with a strand of the fMRI literature over the past decade that has characterised precuneus as a language-relevant region.

6.1.2. Incohesive MWEs and Broca’s Area

While highly cohesive expressions evoke a focal pattern of activation that is distinct from classical left-hemisphere language areas, the cortical sites that showed a sensitivity to a decrease in MWEs’ cohesion strength include the left superior frontal gyrus and also encompass several areas of the perisylvian language network. Notably the left inferior frontal gyrus, or Broca’s area (pars triangularis and pars orbitalis), and also anterior and posterior regions of the left temporal gyrus. These perisylvian areas have been connected with compositional processes (e.g. Pallier et al., 2011), and this is consistent with the leading idea that less-cohesive expressions are more likely to require compositional processes. Indeed a similar set of regions are activated in response to the Bottom-up parsing predictor, as discussed further in the next sub-section.

6.2. Parser Action Count as Composition

Word-by-word composition effort, quantified in terms of bottom-up parser actions, correlates in Analysis 1 with a highly bilateral pattern across several areas in the language network. Notably, both anterior-frontal and anterior-temporal regions are involved bilaterally.
6.2.1. Anterior Frontal and Anterior Temporal regions

Inferior frontal regions encompassing different sub-parts of the inferior frontal gyrus and anterior insula are commonly attributed a role in compositional processes (Friederici & Gierhan, 2013; Hagoort, 2016; Snijders et al., 2009; Zaccarella & Friederici, 2015), and our findings are consistent with that work. Interestingly, prior work using methods most similar to ours (bottom-up parsing, a naturalistic stimulus) have not highlighted the inferior frontal gyrus (Brennan et al., 2012, 2016). We hesitate to draw strong inferences from comparing “significant” to “non-significant” results, but this discrepancy may reflect on the increase in statistical power afforded by our current study, which uses a stimulus that is over seven times longer that used by Brennan and colleagues, and also reports data from almost twice as many participants.

The balance of activation that we observe in response to increasing parser action count is consistent with Ullman’s Declarative-Procedural model. It shows a predominately anterior distribution in left hemisphere, alongside smaller activation clusters in the inferior parietal lobe and temporo-occipital-parietal junction. Further, the observed activation of putamen within the large bilateral cluster in inferior frontal gyrus and anterior insula conforms to Ullman’s prediction that rule-governed language use is instantiated in a Fronto-basal ganglia network (Ullman, 2004, 2015).

The composition effect observed bilaterally in anterior temporal lobe is consistent with previous work on compositional in naturalistic narrative (Brennan et al., 2012). It confirms and extends a broad range of prior findings, including those from simple two-word phrases (Bemis & Pylkkänen, 2011), from parametric variation of constituent size (Pallier et al., 2011), and comparisons of simple sentences to unstructured lists of words (Humphries, Binder, Medler, & Liebenthal, 2006; Rogalsky & Hickok, 2009; Stowe et al., 1998; Vandenberghe, Nobre, & Price, 2002; Xu, Kemeny, Park, Frattali, & Braun, 2005). Our result also conforms with evidence that anterior temporal lesions impair “basic levels of constituent-structure processing” (Dronkers et al., 2004, p. 161).

Recent neuroimaging work correlating brain activity with processing-complexity metrics across grammar types and parsing strategies has convergently pointed to the anterior temporal complex (Brennan et al., 2016; Li & Hale, To Appear) for a review see Brennan, 2016). The precision of this localisation has also been supported by results using intracranial recording (Nelson et al., 2017).

The results reported here, along with previous findings, underline the involvement of anterior temporal lobe in basic composition processes (Friederici & Gierhan, 2013; Hickok & Poeppel, 2007). Yet, along-side many of the findings just reviewed, our results stand in contrast to the apparent stability of compositional processes following anterior temporal atrophy (e.g. Wilson et al., 2014a, 2016). The present study does not resolve the tension between these two literatures. We speculate that such a resolution might involve more detailed consideration of the specific mechanisms that contribute to observed neural activation patterns, and perhaps how compensatory processes respond to neuronal atrophy.

The unexpected right lateralisation of posterior temporal activation may suggest that the way the human brain processes linguistic stimuli within a contextually rich setting — one more similar to the everyday language environment — shows a strongly bilateral involvement of the language network. Future work that parametrically varies contextual richness, from more isolated to more naturalistic stimuli, may shed light on this speculation.

6.2.2. Left Posterior Temporal regions and Parser Action Count

The operationalisation of compositional processing in terms of bottom-up parser actions highlights the language network to a large extent. However, we did not find a significant correlation between this metric and activity in left posterior temporal areas. Our result contrasts with posterior temporal effects of composition that are reported in some prior imaging studies (e.g. Brennan et al., 2016; Pallier et al., 2011; Snijders et al., 2009; Vandenberghe et al., 2002) and are also highlighted in studies of primary progressive aphasia (Wilson et al., 2012). This absence
of evidence should be treated carefully; caution is especially warranted as different studies take different perspectives on composition itself.

The phrase-structures used in our analysis only capture some aspects of composition, and so a more comprehensive view of this aspect of language processing may shed light on discrepancies between studies. For example, the LPTL composition effect reported by Brennan et al. (2016) did not appear when analyzing the same types of tree structures used in the present analysis, but only from tree structures that explicitly encode long-distance dependencies and other more abstract aspects of grammar. More broadly, one can speculate that the absence of activation in our study might be a consequence of the unlexicalised character of the bottom-up parsing metric. It counts distances, for instance along the spine of a tree as in Figure 3 (C), but it treats all nodes as equal, ignoring lexical information except at the preterminal level. Subjects and direct objects receive the same score if they close off the same number of constituents. This conception of composition leaves out many classical distinctions that are known to affect cerebral activity, such as semantic selection, argument structure (Frankland & Greene, 2015; Thompson et al., 2007), long-distance dependencies (Ben-Shachar, Palti, & Grodzinsky, 2004; Santi & Grodzinsky, 2010; Shetreet & Friedmann, 2014), and case-checking (Nieuwland, Martin, & Carreiras, 2012).

Such a more comprehensive perspective may also shed light on another aspect of the present results. While parser action count does not correlate with left posterior temporal activation, we do see such a correlation for decreasing MWE cohesion (see Section 6.1.2). Such a correlation is consistent with the hypothesis that less coherent expressions load more heavily onto certain compositional operations, and such compositional operations evoke, at least in some cases, posterior temporal activity.

Overall the results support the idea that composition and retrieval of stored linguistic elements like MWEs involve different parts of the language network. They also confirm that bilateral anterior frontal regions play an important role in linguistic composition during naturalistic spoken story comprehension.

6.2.3. Composition and the Cerebellum

An often reported but less-discussed area in language processing is the Cerebellum. In this study its relationship to parser action count in Crus I suggests its involvement in compositional processes bilaterally. A growing number of studies have provided evidence for cerebellar involvement in non-motor language functions (see Stoodley & Schmahmann, 2009 for a meta-analysis), revealing interesting patterns of co-activation with classical left prefrontal language areas and right cerebellar hemisphere during language tasks and comprehension (Fedorenko, Duncan, & Kanwisher, 2012), or during the processing of complex syntactic structures (Christensen, 2008; Fabre, 2017; Shetreet & Friedmann, 2014). Specifically, Crus I and Lobule VII are typically proposed as members of Prefrontal-cerebellar connectivity loops (Stoodley & Schmahmann, 2009).

Moreover, an emerging view of perceptual and cognitive processing in the Cerebellum puts special emphasis on sequencing and predictive processes (Leggio et al., 2008; Molinari et al., 2008), which are presumably also required in syntactic sentence parsing operations. Although no special modelling of predictive processing was undertaken in this study, bottom-up approaches such as the LR parser do include predictive information in their control state (see e.g. Shieber & Johnson, 1993; Stabler, 1991).

6.2.4. Composition of Form and of Meaning

The foregoing discussion has remained neutral as to whether activation within the anterior temporal lobe and other brain areas that correlate with bottom-up parsing reflects syntactic processing, or compositional semantic processing. Syntactic structure-building is transparently associated with bottom-up parser action count (panel B of Figure 5) that is itself defined in terms of syntactic phrases such as NP, VP and S. However, our approach is not able to rule
out a semantic explanation. A semantic role would follow within approaches to natural language grammar that adopt the “rule-to-rule” assumption (Bach, 1976). Following the seminal work of Montague (1974) these theories of linguistic competence suppose that syntactic rules generally are paired with corresponding semantic rules. This organisation of grammar, which is introduced briefly in Barker and Jacobson (2007), typifies Combinatorial Category Grammar (Steedman, 2000), Generalized/Head-driven Phrase-Structure Grammars (E. Klein & Sag, 1985) and certain forms of Minimalism (see e.g. Kobele, 2006 §2.2). A consensus from these conceptions of grammar is that aspects of linguistic performance that correlate with syntactic processing will necessarily also correlate with semantic processing.

Of course, the particular rules being processed do matter, and in this case the rules are plainly syntactic in nature. To tease apart semantic from syntactic explanations, future work could develop an alternative grammar, unlike the Penn Treebank phrase structures used here, and evaluate whether this alternative semantically-based theory offers a better explanation of the observed fMRI timecourses. Such an approach may be particularly fruitful in constructions where an additional semantic rule of “coercion” applies, unaccompanied by a corresponding syntactic rule (Pylkkänen, 2008). These considerations, regarding the kinds of rules that are used during comprehension, are in any event quite orthogonal to questions about their order application (Hale, 2014).

7. Conclusion

Analysis of MWEs and parser action counts in naturalistic spoken story comprehension supports the localisation of memory retrieval to the precuneus in a way that trades-off with the other well-known language regions such as inferior frontal gyrus and anterior temporal lobe. This trade-off seems to reflect the degree to which particular MWE form a cohesive unit. These findings are broadly consistent with several contemporary neurocognitive models of language processing, such as the proposals of Hagoort (2016), Friederici and Gierhan (2013) and Ullman (2015).

Apart from the light that they shed on the neural bases of these two language-related cognitive processes, these results also demonstrate the benefits of using combining computational methods for the automatic annotation of narrative texts. Such methods can be leveraged to tease out distinct sub-processes of complex cognitive processes, like language understanding, using experimental setups with greater ecological validity. As such they pave the way for increasing synergy between computational linguistics and the cognitive neuroscience of language.

Acknowledgements

The authors would like to thank Adam Mahar, Skyler Yeatman, Jaelyn Moore, Jacob Collard, Katey Huddleston, and Marissa Holl at Cornell University for assisting us with data collection.

Disclosure statement

We have reported all measures, conditions and data exclusions. The sample size was enlarged by a factor of 2 compared to previous work with similar regressors.

Funding

This material is based upon work supported by the National Science Foundation under Grant Number 1607441 (USA). We also gratefully acknowledge support from the French National Research Agency (ANR) under grant ANR-14-CERA-0001.
References


bridge University Press.


Appendix A. Unthresholded maps

Statistical maps are available at https://neurovault.org/collections/BXCZVBRS/

Appendix B. Correlation matrices

Regressors’ correlation matrices for the models in Analysis 1 and 2 are reported below:

Figure A1. Correlation matrix (Pearson’s $r$) of the convolved regressors included in the GLM model reported in Analysis 1.

Figure A2. Correlation matrix (Pearson’s $r$) of the convolved regressors included in the GLM model reported in Analysis 2.
### Appendix C. Full cluster list for bottom-up parsing regressor

Significant clusters for Bottom-up parser action count (abbreviated table provided in [5.1]):

<table>
<thead>
<tr>
<th>Regions for Bottom-up Parser Action Count</th>
<th>Cluster size (in voxels)</th>
<th>MNI Coordinates</th>
<th>p-value (corrected)</th>
<th>T-score (peak level)</th>
</tr>
</thead>
<tbody>
<tr>
<td>R Anterior Temporal</td>
<td>4816</td>
<td>52 6 -20</td>
<td>0.000</td>
<td>13.20</td>
</tr>
<tr>
<td>R Middle Temporal Gyrus</td>
<td>50</td>
<td>-20 -10</td>
<td>0.000</td>
<td>11.31</td>
</tr>
<tr>
<td>R Supramarginal Gyrus/Superior Temporal Gyrus</td>
<td>60</td>
<td>-40 -10</td>
<td>0.000</td>
<td>10.11</td>
</tr>
<tr>
<td>L IFG Orbitalis/Triangularis (BA47) &amp; Anterior Insula</td>
<td>2461</td>
<td>36 18 -14</td>
<td>0.000</td>
<td>10.40</td>
</tr>
<tr>
<td>L Temporal Pole</td>
<td>-50</td>
<td>6 -26</td>
<td>0.000</td>
<td>8.30</td>
</tr>
<tr>
<td>L Putamen</td>
<td>-30</td>
<td>8 -4</td>
<td>0.000</td>
<td>6.99</td>
</tr>
<tr>
<td>R Supplementary Motor Area/Superior Frontal Gyrus (BA9)</td>
<td>6495</td>
<td>10 18 62</td>
<td>0.000</td>
<td>9.35</td>
</tr>
<tr>
<td>L Medial Superior Frontal Gyrus (BA9)</td>
<td>12</td>
<td>58 32</td>
<td>0.000</td>
<td>8.62</td>
</tr>
<tr>
<td>L Superior Frontal Gyrus</td>
<td>-8</td>
<td>18 66</td>
<td>0.000</td>
<td>8.24</td>
</tr>
<tr>
<td>L Cerebellum – Crus I/II</td>
<td>448</td>
<td>-24 -74 -30</td>
<td>0.000</td>
<td>8.96</td>
</tr>
<tr>
<td>R Cerebellum – Crus I/II</td>
<td>941</td>
<td>26 -74 -36</td>
<td>0.000</td>
<td>8.15</td>
</tr>
<tr>
<td>L IFG Orbitalis/Triangularis (BA47) &amp; Anterior Insula</td>
<td>2461</td>
<td>36 -60 -32</td>
<td>0.021</td>
<td>5.16</td>
</tr>
<tr>
<td>L Temporal Pole</td>
<td>-50</td>
<td>6 -26</td>
<td>0.000</td>
<td>8.30</td>
</tr>
<tr>
<td>L Putamen</td>
<td>-30</td>
<td>8 -4</td>
<td>0.000</td>
<td>6.99</td>
</tr>
<tr>
<td>R Supplementary Motor Area/Superior Frontal Gyrus (BA9)</td>
<td>6495</td>
<td>10 18 62</td>
<td>0.000</td>
<td>9.35</td>
</tr>
<tr>
<td>L Medial Superior Frontal Gyrus (BA9)</td>
<td>12</td>
<td>58 32</td>
<td>0.000</td>
<td>8.62</td>
</tr>
<tr>
<td>L Superior Frontal Gyrus</td>
<td>-8</td>
<td>18 66</td>
<td>0.000</td>
<td>8.24</td>
</tr>
<tr>
<td>L Cerebellum – Crus I/II</td>
<td>448</td>
<td>-24 -74 -30</td>
<td>0.000</td>
<td>8.96</td>
</tr>
<tr>
<td>R Cerebellum – Crus I/II</td>
<td>941</td>
<td>26 -74 -36</td>
<td>0.000</td>
<td>8.15</td>
</tr>
<tr>
<td>L IFG Orbitalis/Triangularis (BA47) &amp; Anterior Insula</td>
<td>2461</td>
<td>36 -60 -32</td>
<td>0.021</td>
<td>5.16</td>
</tr>
<tr>
<td>L Temporal Pole</td>
<td>-50</td>
<td>6 -26</td>
<td>0.000</td>
<td>8.30</td>
</tr>
<tr>
<td>L Putamen</td>
<td>-30</td>
<td>8 -4</td>
<td>0.000</td>
<td>6.99</td>
</tr>
</tbody>
</table>

---

29