Grounding the cognitive neuroscience of semantics in linguistic theory

Liina Pylkkänen, Jonathan Brennan, and Douglas K. Bemis

Departments of Linguistics and Psychology, New York University, New York, NY, USA

The mission of cognitive neuroscience is to represent the interaction of cognitive science and neuroscience: cognitive models of the mind guide a neuroscientific investigation of the brain bases of mental processes. In this endeavour, a cognitive model is crucial as without it, the cognitive neuroscientist does not know what to look for in the brain, what the nature of the relevant representations might be, or how the different components of a process might interact with each other. In the cognitive neuroscience of language, the interaction of theoretical models and brain research has, however, been far from ideal, especially when it comes to the study of meaning at the sentence level. Although theoretical semantics has a long history in linguistics and thus offers detailed and comprehensive models of the nature of semantic representations, these theories have had minimal impact on the brain investigation of semantic processing. In this article, we outline what a theoretically grounded cognitive neuroscience of semantics might look like and summarise our own findings regarding the neural bases of semantic composition, the basic combinatorial operation that builds the complex meanings of natural language.

Keywords: Formal semantics; Cognitive neuroscience; Magnetoencephalography; AMF.

If a visitor from Mars was told what the cognitive neuroscience of language is and what linguistics is, they would presumably immediately assume that the two disciplines must be tightly connected—after all, how could one study the neural bases of language without an understanding of what language is? At
the present time, however, the connection between the two fields remains rather loose. Linguists have not done a terrific job in making their results accessible to the general public, and perhaps neuroscientists have also not reached out sufficiently to seek theoretical grounding for their research. In our work, we seek to improve the situation for one particular subdomain of language: combinatory semantics.

Of the various core processing levels of language, i.e., phonetics, phonology, syntax, and semantics, the cognitive neuroscience of semantics has been the most divorced from linguistics. For example, for phonology, although there is no cognitive neuroscience for most questions that phonologists are interested in, there is, for instance, a wide literature studying the neural bases of phonological categories assuming exactly the types of phoneme representations proposed by phonological theory (e.g., Dehaene-Lambertz, 1997; Hickok & Poeppel, 2000; Nääätänen et al., 1997). Similarly for syntax, research probing into the function of Broca’s area, for example, has assumed tree representations of the sort that one would also find in an introductory syntax textbook (e.g., Caplan, Alpert, & Waters, 1999; Embick, Marantz, Miyashita, O’Neil, & Sakai, 2000; Grodzinsky & Friederici, 2006; Moro et al., 2001; Patel, 2003; Stromswold, Caplan, Alpert, & Rauch, 1996). Not so for semantics. In fact, the word “semantics” tends to mean rather different things in cognitive neuroscience and in linguistics. In linguistics, the semantics of an expression refers to the complex meaning that is computed by combining the meanings of the individual lexical items within the expression. The formal rules that are employed in this computation have been studied for about 40 years, and many aspects of these rules are now well-understood. But until recently, despite the clear necessity of understanding the neurobiology of the combinatory functions that derive sentence meanings, we have had no cognitive neuroscience on the brain bases of these rules. Instead, cognitive neuroscience research with “semantic” in the title most often deals with the representation of conceptual knowledge, investigating questions such as whether living things and artifacts are distinctly represented in the brain (e.g., Mummery, Patterson, Hodges, & Price, 1998), or how different types of object concepts are represented (Martin & Chao, 2001). The fact that this literature is disconnected from linguistics is not devastating—the questions are rather different from those that concern linguists. Although linguistic semantics does include a subfield called “lexical semantics” that focuses on word meaning, even this subfield is mostly interested in word meaning that appears to be composed of smaller parts—i.e., it has some type of complexity to it. The representational differences between tools and animals are not a big research question in linguistics and thus the brain research on conceptual knowledge has less to gain from linguistic theories.

There is another line of brain research on “semantics”, however, that should and needs to connect with the types of theoretical models of meaning
representation offered by linguistics. This is the so-called “N400 literature”, which uses the N400 as an index of “semantic integration”. Although this functional interpretation is assumed by a sizeable research community (see Lau, Phillips, & Poeppel, 2008 for a recent review), the N400 was not discovered via a systematic search for a neural correlate of a theoretically defined notion of “semantic integration”. In fact, the behaviour and the source localisation of the N400 are much more compatible with a lexical access-based account (Lau et al., 2008), making the N400 an unlikely index of semantic integrative processes.

In this article, we outline what we believe a theoretically grounded cognitive neuroscience of semantics should look like. Our focus is on combinatory semantics, i.e., the composition operations that serve to build complex meanings from smaller parts. We take formal syntax and semantics of the generative tradition (e.g., Heim & Kratzer, 1998) as the cognitive model that guides this research and defines the operations whose neurobiology is to be investigated. We assume, hopefully uncontroversially, that the right way to uncover the neural bases of semantic composition is to systematically vary, or otherwise track, this operation during language processing. In our own research, we have aimed to do exactly this. We will first define exactly what we mean by “semantic composition”, then summarise our findings so far, discuss some open questions, and finally, articulate our hopes for the future.

SEMANTIC COMPOSITION: DEFINING THE CORE OPERATIONS

Theories of formal semantics are models of the possible meanings and semantic combinatorics of natural language (e.g., Dowty, 1979; Heim & Kratzer, 1998; Montague, 1970; Parsons, 1990; Steedman, 1996). They aim to give a complete account of the representations and computations that yield sentence meanings, including relatively straightforward rules that combine predicates with arguments and adjuncts and extending to more complicated phenomena such as the interpretation of tense, aspect, focus, and so forth. The goal of these models is to understand and formalise the nature of semantic knowledge both within languages and cross-linguistically. Consequently, theories of formal semantics provide an extremely rich and detailed hypothesis space both for psycho and neuro-linguistic experimentation.

One challenge for rooting the cognitive neuroscience of semantics in formal theories of meaning is deciding what theory to follow; unsurprisingly, different models make different claims about semantic representations and the mechanisms by which they are computed. Theoretical semantics is a lively field with debate at every level of analysis, ranging from foundational
questions (such as what is the basic relationship between syntax and semantics?) to issues relating to the formal details of the analysis of a specific construction within a specific language. Such variance of opinion may seem bewildering to a neuroscientist trying to figure out what aspects of natural language meaning semanticists agree on, such that those phenomena could safely be subjected to a brain investigation.

Unfortunately, there is no safe route: like any other branch of cognitive science, semantic theory is ever evolving and virtually every aspect of it has been or at least can be questioned. However, it is still possible to identify certain basic operations as a starting point, while keeping in mind the possibly controversial aspects of the formal treatment. In what follows, we describe two operations that form the core of the combinatory engine in most semantic theories (for a fuller exposition, see Pylkkänen & McElree, 2006). Our notation follows Heim and Kratzer (1998), but again, for our present purposes the details of the rules are not important; rather, what is crucial is that this type of theory distinguishes between two different modes of composition, one that serves to fill in argument positions of predicates and one that modifies the predicate without impacting its arguments. Although mainstream, this distinction is neither a formal necessity nor necessarily empirically correct (e.g., Pietroski, 2002, 2004, 2006), but given its central role in most semantic theories, it is one of the most basic distinctions that can be subjected to a brain investigation. Since one reason for the disconnection between cognitive neuroscience and formal semantics has likely been the relative impenetrability of semantic theories for the nonlinguist, in the following we aim for a maximally informal and accessible description of the basic ideas.

Function application

The driving force behind mainstream theories of semantic composition is the idea that the meanings of most words are in some sense incomplete, or “unsaturated”, unless they are combined with other words with suitable meanings (Frege, 1892). For an intuitive example, the semantics of action verbs is thought to have “placeholders”, or variables, that stand for the participants of the actions described by the verbs. In order for the meanings of these verbs to be complete, or saturated, the verb needs to combine with noun phrases that describe those participants. More formally, these verbs are treated as functions that take individual-denoting noun phrases as their arguments. This idea is expressed in lambda calculus for the transitive verb destroy in (1) below. The arguments of the function are prefixed with lambdas, and the value, or the output of the function, follows the lambda terms:
In order to combine the verb *destroy* with its arguments, we apply a rule called *function application*, which essentially replaces the variables in the denotation of the predicate and erases the lambda prefixes. Proper names are the most straightforward individual-denoting noun phrases and thus the representation of the sentence *John destroyed Bill* would involve roughly the following representation (here we ignore everything that does not pertain to the argument saturation of the verbal predicate).

In addition to simple examples such as the one above, function application is the mode of composition for any predicate and its argument. For example, prepositions combine with noun phrases via function application [in [the dark]] as do verbs with clausal complements [knew [that John destroyed Bill]] and verbal ones [John tried [to destroy Bill]]. Thus function application accounts for a vast amount of semantic composition in natural language and every theory of semantics requires a rule that accomplishes its task. Consequently, it should be regarded as one of the core operations that any theory of the brain bases of language should address.

**Predicate modification**

The gist of function application is that it satisfies the semantic requirements of a predicate and makes it more complete. But most expressions also have parts that are not formally required, i.e., they could be dropped and the expression would still be grammatical. For instance, although we could describe The New York Times simply as a *newspaper*, we could also describe it as a *major national newspaper*. There are no grammatical reasons for why we would ever have to add these adjectives—unlike with *destroy* above, which simply could not function grammatically without, say, its object (*John destroyed*). The adjectives simply provide additional, potentially useful information about the noun. The majority of these types of optional elements are added into expressions via a rule called *predicate modification*.
Predicate modification builds complex properties from simpler ones. In formal semantics, adjectives and nouns are both treated as describing properties of individuals, e.g., *newspaper* denotes the property of being a newspaper, and *major* the property of being major. Predicate modification simply concatenates these properties, such that *major national newspaper* ends up meaning “the property of being major, and national, and a newspaper”, as shown in lambda notation in (3).

\[
\lambda x. \text{major}(x) \& \text{national}(x) \& \text{newspaper}(x) \quad \text{[by predicate modification]}
\]

\[
\begin{array}{c}
\text{major} \\
\lambda z. \text{major}(z)
\end{array}

\begin{array}{c}
\text{national} \\
\lambda y. \text{national}(y)
\end{array}

\begin{array}{c}
\text{newspaper} \\
\lambda x. \text{newspaper}(x)
\end{array}

\quad \text{[by predicate modification]}
\]

In addition to building complex properties of individuals, predicate modification can also combine any two predicates that describe properties of the same type of thing. For example, in event-based semantic frameworks (e.g., Davidson, 1967; Kratzer, 1996; Parsons, 1990), both verbs and adverbs are thought of as describing properties of events. Thus the mode of composition between, say, *run*, and *quickly* would be a version of predicate modification, such that the resulting verb phrase would describe events that are both quick and events of running.

In summary, while function application is the default mode for adding arguments into a structure, predicate modification is the most basic way to add optional elements. Although natural language semantics as a whole requires a much more elaborate system than just these two rules, function application and predicate modification constitute the nuts and bolts of semantic composition in most theories, and thus offer clearly defined computations whose neural bases can be characterised within the cognitive neuroscience of language. It is of course an empirical question whether function application and predicate modification dissociate in brain activity. If they do not, this would be a type of null result and not necessarily a motivation for linguists to revise their theories, but a lack of dissociation would obviously call into question the assumption that the operations are distinct. In this type of situation, the brain data would be more compatible with a theory where semantic composition is achieved via a unified operation across expression types, as proposed, for example, by Pietroski (2002, 2004, 2006). In the long run, this is the type of interaction between brain experimentation and linguistic theorising that we advocate.
Theories of formal semantics have a crisp definition of semantic well-formedness: an expression is semantically well-formed if the rules of composition yield a meaning for it. For example, in the default case, there is no way for a verb such as arrive, i.e., a function that takes a single individual-denoting argument, to combine with two such arguments, given that this predicate only has one argument slot for function application to saturate. Thus an expression such as I arrived Helsinki is considered uninterpretable under this type of theory; to make it interpretable, another predicate would need to be added, such as the preposition in, which can then take Helsinki as its argument.

This definition of well-formedness is very different from layman’s intuition of what “makes sense”, i.e., an expression can be semantically well-formed in a formal sense even if it means something crazy. For example, I just ate a cloud is a nonsensical sentence in that the situation it describes could never happen (unless the world was significantly altered). However, every native speaker of English should have the intuition that they know what this sentence means and, in fact, this meaning is easily computed by the two rules introduced above. In terms of formal semantics, the oddness of the sentence is not revealed until one compares its semantic representation (computed by the rules) to one’s knowledge of the world, which dictates that clouds are not edible entities. But crucially, the sentence is not semantically ill-formed, but rather just ill-fitting to our world knowledge.

This conception of semantic well-formedness is dramatically different from what is generally assumed in neuroimaging and especially in event-related potentials (ERP) research on semantics. Such brain research has been overwhelmingly dominated by studies that vary “semantic well-formedness” or “congruity”, but in the majority of these experiments, the semantically ill-formed stimuli are not ill-formed in the above described, formal sense; rather, in terms of linguistic theory, they violate world knowledge instead. The original ERP study on such violations, by Kutas and Hillyard (1980), used sentences such as he spread the warm bread with socks to violate what they called “semantic appropriateness”. These violations elicited a negative-going ERP peaking around 400 ms, the N400, a finding that has since been replicated in hundreds of studies. Since the N400 is elicited by semantically surprising elements, it is commonly thought of as a semantics-related ERP. However, this notion of “semantics” is not theoretically grounded, i.e., we are not aware of any theory of linguistic representation according to which typical N400 sentences would count as semantically ill-formed (and even if there is such as a theory, the N400 literature does not reference it). For instance, in the above example he spread the warm bread with socks, the final
word *socks* composes without a problem with the preposition *with*, to yield a meaning in which socks are used as a spread on bread. Now socks are of course not edible or spreadable, but in terms of linguistic theory, this is a fact about the world, not a fact about language. Given that there is no crisp definition of what “semantics” means in the N400 literature, considering this component a semantics-related ERP is not conducive to building a theoretically grounded model of language in the brain.

It is an empirical question whether N400-type effects could also be obtained for expressions that are semantically ill-formed in the linguistic sense. There are a few studies that have run relevant manipulations, primarily focusing on violations of negative polarity licensing (Drenhaus, beim Graben, Saddy, & Frisch, 2006; Saddy, Drenhaus, & Frisch, 2004; Steinhauer, Drury, Portner, Walenski, & Ullman, 2010; Xiang, Dillon, & Phillips, 2009). Negative polarity items (NPIs) are expressions such as *ever* or *any* that must occur in the scope of negation¹ (e.g., *No tourist ever visited the old harbour* vs. *One tourist ever visited the old harbour*). Violations of this licensing condition count as true semantic violations in terms of linguistic theory, in that if an NPI is not licensed by a suitable operator, no semantic representation can be built for the sentence. Negative polarity violations have indeed been reported to elicit N400 effects (Saddy et al., 2004), but also P600 effects (Drenhaus et al., 2006; Xiang et al., 2009) and late left anterior negativities (Steinhauer et al., 2010). Thus the ERP profile of these violations has been quite variable and does not easily lend itself to firm conclusions. Most relevantly for our present purposes, however, negative polarity violations have exhibited rather different ERP profiles from world knowledge violations when the two are examined within the same study (Steinhauer et al., 2010).

In summary, most extant research on the cognitive neuroscience of sentence level meaning has operated with only an intuitive notion of “semantics”, not grounded in any particular theoretical model. It should, however, be uncontroversial that technical terminology in any field needs to be clearly defined; otherwise it is unclear what the research is even about. Given the vast amount of work devoted to carefully characterising crisp, formal semantic theories, disconnecting the cognitive neuroscience of semantics from these models is a missed opportunity. Only by grounding the neurobiology of meaning in theoretical models of meaning can we move forward to a “predictive mode” of research, where the properties of a cognitive model guide the experimentation and make predictions about what phenomena should group together. Without such grounding, one is left with complex data-sets and a layman’s intuition about what it all may mean.

¹ Or in the scope of some other operator that supports decreasing inferences (Ladusaw, 1979, 1980, 1983).
THE CHALLENGE OF COMPOSITIONALITY

As soon as one couches the brain investigation of semantics in linguistic theory, a methodological dilemma, however, arises: In most cases, the semantic complexity of an expression tightly correlates with its syntactic complexity, and thus it is not obvious how one could ever vary semantic composition without syntactic confounds. Even worse, linguistic theories differ in the extent to which syntax and semantics are coupled. In some theories, every step of the semantic derivation corresponds to a step in the syntactic derivation (e.g., Heim & Kratzer, 1998; Montague, 1970). Other theories assume a weaker coupling between syntax and semantics, allowing certain operations to only apply in the semantics without a corresponding syntactic step (Barker, 2002; Hendriks, 1988; Jacobson, 1999; Partee & Rooth, 1983). And finally, in Jackendoff’s parallel model, syntactic and semantic representations are built by completely independent mechanisms, creating an architecture that forces no correlation between the number of syntactic and semantic steps (Jackendoff, 2002). The disagreement between these theories has to do with the degree of compositionality in natural language, i.e., the extent to which the meanings of expressions are straightforwardly determined by the meanings of their parts and the syntactic combinatorics.

With this type of uncertainty about the fundamental relationship between syntax and semantics, how could a cognitive neuroscientist use these theories to make headway in understanding the brain bases of semantic composition? In our own research, we have adopted what might be called a “bite the bullet” approach. First, we have to accept that there are no expressions that uncontroversially involve semantic computations that do not correspond to any part of the syntax: even if some meaning component of an expression does not appear to map onto the syntax, one can always postulate a phonologically null syntactic element that carries that piece of meaning. The question though is, whether such an analysis makes the right empirical predictions. As discussed in Pylkkänen (2008), there is great asymmetry in the degree to which this type of solution works for different cases; it sometimes makes the right predictions, but often not. As a starting point in our research, we have aimed to study exactly the expressions that are most likely to involve syntactically unexpressed semantic operations, even if this dissociation cannot be definitively proven. As discussed in the next section, our subsequent research has shown that these types of cases pattern differently from cases where syntax and semantics are uncontroversially coupled, lending support to the notion that the initial set of test cases did, in

2 Although such a correlation can and must be obviously be built in, given the descriptive generalisation that syntactic and semantic operations do, by and large, stand in a one-to-one correspondence.
fact, only vary semantics. Thus compositional coupling of syntax and semantics makes the project of “neurosemantics” a difficult one to get off the ground, but once some initial progress is made, we believe it may be possible for brain data to actually contribute to debates about the relationship between syntax and semantics.

THE ANTERIOR MIDLINE FIELD (AMF) AS A CORRELATE OF SEMANTIC COMPOSITION

Varying semantic composition: semantic mismatch

Although much of natural language appears strongly compositional, there are classes of expressions whose meanings appear richer than their syntax. Perhaps the best studied example is so-called complement coercion, illustrated in (4) below.

(4) a. The professor began the article.
    b. The boy finished the pizza.
    c. The trekkers survived the mountain.

When native speakers are queried about the meanings of these sentences, they typically report for (4a) that the professor began reading the article, (4b) that the boy finished eating the pizza, and (4c) that the trekkers survived climbing the mountain. Reading, eating, or climbing do not, however, figure in the lexical material of these expressions. Thus where do these implicit activity senses come from? They are typically thought of as arising from a certain semantic mismatch between the verbs and their direct objects. Semantically, each of the verbs in (4) selects for a direct object that describes some type of event or activity: one generally begins, finishes, or survives doing something. However, none of the direct objects in (4) denote entities. This kind of situation is formally considered a semantic “type mismatch” and in the default case, type mismatch leads to ungrammaticality. However, the sentences in (4) are all grammatical and thus the type mismatch must be somehow resolved. Descriptively, the resolution appears to involve the semantic insertion of an implicit activity that can mediate between the verb and the otherwise ill-fitting object NP. Formally, the complement of the verb (the object NP) is thought to “coerce” into an event meaning (Jackendoff, 1997; Pustejovský, 1995), such that semantic composition can succeed. This type of analysis treats coercion as a purely semantic meaning-adding operation, with no consequences for the syntax.3

3 For an extensive discussion on the possible empirical arguments for this, see Pylkkänen (2008).
Coercion does, however, have consequences for online processing: numerous studies have shown that coercion expressions take longer to process than control expressions involving no coercion (for a review, see Pylkkänen & McElree, 2006). This psycholinguistic evidence provides empirical support for the hypothesis that coercion expressions do, in fact, involve some type of extra computation, absent in more transparently compositional expressions. Given that the processing of complement coercion had already been well-studied behaviourally, we used it as the initial test case in our attempt to identify a neural correlate of purely semantic combinatoric processing.

To measure brain activity, our research uses magnetoencephalography (MEG), which offers the best combination of both spatial and temporal resolution of existing cognitive neuroscience techniques. MEG measures the magnetic fields generated by neuronal currents. The primary difference between MEG and EEG is that in MEG, the locations of the current generators can be estimated reasonably accurately given that magnetic fields pass through the different structures of the brain undistorted, contrary to electric potentials. Typically, the current sources of MEG recordings as modelled either as focal sources (so-called single dipoles) or as distributed patches of activation on the cortex (typically minimum norm estimates) (Hansen, Kringelbach, & Salmelin, 2010). The intensities and latencies of these current sources then function as the primary dependant measures of most MEG studies, although an EEG-style analysis of sensor data without source localisation is also always an option.

When we contrasted coercion expressions (the journalist began the article) with noncoercive control sentences (the journalist wrote the article) during an MEG recording, we observed increased activity for coercion in a prefrontal midline MEG field, dubbed the anterior midline field (AMF) (Figure 1). Source localisation indicated the ventromedial prefrontal cortex (vmPFC) as the generating brain region of this magnetic field. No such effect was observed for implausible control sentences, suggesting different brain bases for the computation of coercion and the detection of poor real-world fit.

These complement coercion findings established a starting point for our research on the neural bases of semantic composition. Our subsequent studies have then aimed to narrow down the possible functional interpretations of this activity. Figure 1 summarises all of our AMF results so far. First, we examined whether the AMF effect generalises to other coercion constructions, and found that it is also observed for a different variant of complement coercion (Pylkkänen, Martin, McElree, & Smart, 2009) as well as for two different types of aspectual coercion (Brennan & Pylkkänen, 2008, 2010). These findings ruled out the hypothesis that the AMF reflects processing specific to complement coercion.

We have also examined the AMF in a violation paradigm (Pylkkänen, Oliveri, & Smart, 2009), to better connect our findings to ERP research.
which has been dominated by this type of design. In this study, we again used semantic mismatch, but of a sort that cannot be resolved by a productive rule, but rather results in semantic ill-formedness (in the linguistic sense, i.e., no well-formed representation can be built). These semantic violations were contrasted both with violations of world knowledge (similar to the “semantic violations” of the N400 literature) and with well-formed control expressions,
in order to assess in an ERP-style violation paradigm whether semantic violations but not world knowledge violations would affect the AMF, as would be predicted if this activity reflected the composition of semantic representations but not the evaluation of their real-world fit. Our results indeed patterned according to this prediction: semantic violations elicited increased AMF amplitudes, while world knowledge violations generated a different type of effect.

Thus our initial set of experiments employed resolvable and unresolvable semantic mismatch in order to vary semantic composition while keeping syntax maximally constant. These studies yielded highly consistent results, implicating the AMF MEG component as potentially relevant for the construction of complex meaning. This effect demonstrated task-independence (Pylkkänen, Martin, et al., 2009) and was insensitive to world knowledge (Pylkkänen, Oliveri, et al., 2009). But of course these results do not yet show that we have isolated activity reflecting semantic composition in some general sense, as opposed to activity that plays a role in mismatch resolution but does not participate in more ordinary composition. Furthermore, the psychological nature of coercion operations is vastly underdetermined by traditional linguistic data (e.g., judgements of grammaticality and truth value); in some sense “coercion” and “type-shifting” are descriptive labels for computations that matter for interpretation but do not appear to have a syntactic nature. In other words, although coercion rules are traditionally thought of as operating within the compositional semantic system, it is difficult to demonstrate this empirically, i.e., for example, to rule out the possibility that these meaning shifts might be essentially pragmatic in nature. Having discovered that coercion affects a particular brain response, i.e., the AMF, it becomes possible to investigate whether the same response would also be affected by simpler manipulations of semantic composition, involving no mismatch resolution. If ordinary “run-of-the-mill” composition was to also modulate the AMF, this would show that the role of the AMF is not limited to mismatch resolution and would also offer empirical support for the hypothesis that coercion operations involve semantic computations similar to those involved in transparently compositional expressions. The research summarised below aimed to assess this by studying very simple cases of composition, involving only the intersective combination of a noun and an adjective.

Simple composition: bringing in syntax and the left anterior temporal lobe (ATL)

Our study on simple composition constitutes, to our knowledge, the first neurolinguistic investigation directly targeting one of the core semantic operations outlined in Section “Semantic composition: defining the core
operations” above. Specifically, we aimed to measure the MEG activity associated with predicate modification (Bemis & Pylkkänen, 2010). Contrary to most extant brain research on syntax and semantics, which has generally employed full sentences involving complex structures such as centre embedding or semantic anomalies (for reviews, see Kaan & Swaab, 2002; Lau et al., 2008), our study employed minimal phrases invoking exactly one step of semantic composition. The critical stimulus was an object-denoting noun that was either preceded by a semantically composable colour adjective (red boat) or by consonant string activating no lexical meaning (xkq boat). Subjects viewed these phrases (and other control stimuli) and then decided whether or not a subsequent picture matched the verbal stimulus. If the role of the AMF is limited to mismatch resolution, it clearly should not be affected by this maximally simple manipulation. In contrast, if the AMF reflects aspects of semantic composition quite generally, nouns preceded by adjectives should elicit increased AMF activity. The latter prediction was robustly confirmed: a clear AMF amplitude increase was observed for the adjective–noun combinations relative to the isolated nouns, ruling out the hypothesis that the AMF is purely mismatch related.

But predicate modification is of course not the only combinatory operation varied in this manipulation: each adjective–noun pair also made up a syntactic phrase. Thus the above contrast should also elicit effects related to the syntactic side of the composition, if this is indeed something separable from the semantic combinatorial effects. We did, in fact, also observe a second effect for the complex phrases, and in a location familiar from a series of prior studies. This effect localised in the left anterior temporal lobe (ATL), which has been shown by several imaging studies as eliciting increased activation for sentences as opposed to unstructured lists of words (Friederici, Meyer, & von Cramon, 2000; Humphries, Binder, Medler, & Liebenthal, 2006; Mazoyer et al., 1993; Rogalsky & Hickok, 2008; Stowe et al., 1998; Vandenberghe, Nobre, & Price, 2002). This body of work has hypothesised that the left ATL is the locus of basic syntactic composition, with Broca’s region only participating in more complex operations. This interpretation is further corroborated by imaging results of our own, showing that while subjects listen to a narrative, activity in the left ATL correlates with the number of syntactic constituents completed by each word (Brennan et al., 2010). Thus the combination of the results reviewed so far suggests the following working hypothesis: the AMF generator, i.e., the vmPFC, and the left ATL are the primary loci of basic combinatorial processing, with the vmPFC computing semantic and the left ATL syntactic structure.

The hypothesis just articulated raises a puzzle, though, regarding the relationship between our results on the vmPFC and the just mentioned imaging studies contrasting sentences vs. word lists. Clearly, the sentence vs.
word list contrast varies not only syntactic composition but also semantic composition, yet none of the studies using this contrast have reported effects in the vmPFC. This is obviously incompatible with our semantic composition hypothesis regarding this region. But the sentence vs. word list studies have also used a different technique from our research, measuring slowly arising hemodynamic activity as opposed to electromagnetic activity, which can be measured at a millisecond temporal resolution, matching the speed of language processing. To assess whether a vmPFC effect for sentences over word lists would be observed in MEG, we conducted an MEG version of the traditional imaging design (Brennan & Pylkkänen, 2010). Our results indeed revealed an amplitude increase in the vmPFC for sentences, conforming to the composition hypothesis. An effect was also seen in the left ATL, replicating the prior imaging results. While questions remain about why hemodynamic methods and MEG should yield different results here, the picture emerging from our MEG findings is encouraging and begins to have the shape of an elementary theory about the neural bases of syntactic and semantic composition, grounded in linguistic theory.

**Challenges and open questions**

*Relation to hemodynamic and neuropsychological data*

Our MEG research so far suggests a prominent role for the generating region of the AMF in language processing. However, if the role of this midline prefrontal region is as basic as semantic composition, how come this area has not already figured in neuroimaging studies on sentence processing? As just discussed, it is possible that MEG may be better suited for capturing this activity than hemodynamic methods, given our findings on the sentence vs. word list paradigm which yields a vmPFC effect in MEG but not in fMRI or PET (granted that we have not yet conducted a study with the same materials and subjects across the techniques). One possible reason for this difference lies in the better time resolution of MEG: a short-lived time-locked effect lasting less than 100 ms should naturally be difficult to capture with techniques whose temporal resolution is on the scale of several seconds. Furthermore, fMRI in particular is limited when it comes to measuring signal from the vmPFC and the orbitofrontal cortex in general. This is due to the proximity of these regions to the sinuses, which leads to so-called susceptibility artifacts, resulting in significant signal loss in orbitofrontal regions (Ojemann et al., 1997). Despite these limitations, however, several hemodynamic studies have shown interpretation-related effects in the vmPFC, both in PET (e.g., Maguire, Frith, & Morris, 1999; Nathaniel-James & Frith, 2002) and in fMRI (Nieuwland, Petersson, & Van Berkum, 2007). Thus it is not the case that our findings are entirely without precedent in the neuroimaging literature.
One type of evidence that does not suffer from the artifact issues described above pertains to neuropsychological data. Do patients with ventromedial prefrontal damage suffer severe language processing deficits, of the sort one would expect of a person who has lost their ability to construct complex linguistic meanings? The most simple-minded prediction of the semantic composition hypothesis might be that such patients should not be able to understand anything or be able to produce any coherent sentences. This prediction is certainly not born out: language problems are not the predominant issue for this patient population. Instead, their most salient impairments have to do with appropriateness of social behaviour, decision-making, and emotion regulation (Anderson, Barrash, Bechara, & Tranel, 2006; Barrash, Tranel, & Anderson, 2000; Berlin, Rolls, & Kischka, 2004; Burgess & Wood, 1990; Grafman et al., 1996; Koenigs & Tranel, 2007). Perhaps because language problems do not present prominently, language skills are typically not even reported, or they may be informally commented on as “normal” (e.g., Anderson et al., 2006). There is though one body of research on vmPFC patients that has focused on language processing, specifically on sarcasm and irony. In this work, vmPFC damage has been found to lead to deficits in comprehending sarcasm, a finding that has been related to the role of the vmPFC in theory-of-mind processing (Shamay-Tsoory, Tibi-Elhanany, & Aharon-Peretz, 2006; Shamay-Tsoory, Tomer, & Aharon-Peretz, 2005). This finding is obviously related to semantic processing, but not at the very basic level that our most recent simple composition results suggest (Section “Simple composition: bringing in syntax and the left anterior temporal lobe (ATL)”).

What then, to make of these relatively sparse neurolinguistic data on vmPFC patients in light of the semantic composition hypothesis of the AMF? At least two considerations are important to note here, one methodological and the other theoretical. The methodological consideration is that although MEG systematically localises the AMF in the vmPFC (with some variation on the anterior–posterior axis), this localisation is a source model and not necessarily the absolute truth. For example, so far most of our source localisations have used a smoothed average cortex (provided by BESA, the source localisation software), which does not include the medial wall. In this kind of source model, activity in the anterior cingulate, for example, might be projected onto vmPFC. Thus it is not yet obvious that vmPFC patients are necessarily the right clinical population to consider. Clearly, the localisation of the AMF will need to be further assessed with hemodynamic methods, although as just reviewed, this approach presents challenges of its own.

The theoretical consideration has to do with the precise predictions of the semantic composition hypothesis. What should a person’s linguistic performance look like if they have lost their ability of semantic composition
but have intact syntax and lexical-semantic processing? The answer is far from obvious. Such a person would have a combinatorial system, i.e., the syntax, and they would understand the meanings of words. Thus it is possible that they might be able to reasonably work out what sentences mean, only failing in subtler situations. Thus perhaps the only clear prediction of the semantic composition hypothesis is that a person without the relevant region should show some semantic deficits. Whether vmPFC patients fit this profile is currently unclear; assessing this would require sophisticated linguistic testing perhaps focusing on semantic phenomena that are not transparently reflected in the syntax.

**Domain generality**

One of the most fundamental questions regarding the neural architecture of language is the extent to which the mechanisms employed by language are specific to language or also used in other domains. Localisation wise, our findings on the possible brain bases of semantic composition have been quite surprising: the midline prefrontal regions that are modelled as the AMF source are not generally thought of as “language regions”. Instead, these areas, and the vmPFC specifically, are prominently associated with various types of nonlinguistic higher cognition, such as emotion (Bechara, Damasio, & Damasio, 2000; Damasio, 1994), decision-making (Wallis, 2007), representation of reward value (Schoenbaum, Roesch, & Stalnaker, 2006), and social cognition, including theory-of-mind (Amodio & Frith, 2006; Baron-Cohen & Ring, 1994; Gallaher & Frith, 2003; Krueger, Barbey, & Grafman, 2009; Rowe, Bullock, Polkey, & Morris, 2001). Thus it seems plausible that the AMF may reflect computations that span across multiple domains, a speculation we already put forth in the first AMF paper (Pylkkanen & McElree, 2007) and have later refined (Pylkkanen, 2008; Pylkkänen, Oliveri, et al., 2009). A more specific way to put this hypothesis is that the AMF reflects semantic composition and that semantic composition employs mechanisms also used in other domains. To test this, our on-going research is aimed at assessing whether the AMF composition effects extend to combinatorial phenomena in nonlinguistic domains.

**LOOKING AHEAD**

Moving forward, here is what we would like to not see happen in the cognitive neuroscience of semantics: for neuroscientists to “reinvent the wheel” when it comes to cognitive models of semantic representation. We understand that linguistic theory offers a tremendous amount of detail, which may seem daunting to try to relate to neurobiological data. Thus it may be tempting to flatten some of those distinctions when aiming to characterise the brain bases
of linguistic functions, as in, for example, Peter Hagoort’s model, where all combinatory operations of language are treated as examples of a general “unification” function (Hagoort, 2005). We actually think it is quite possible that the brain performs composition in various domains in rather similar ways—this intuition is largely what drives our work on the domain generality of the AMF—but we do not believe that a flat cognitive model is the right kind of theory to guide us when designing experiments. If we do not look for certain (theoretically well-motivated) distinctions in the brain, we certainly will not find them, should they, in fact, be there.

To conclude, natural language meaning exhibits many fascinating phenomena, most of which one is not consciously aware of without engaging in the formal study of semantics. If theoretical semantics and the cognitive neuroscience of semantics are not in communication with each other, we will never understand how the brain achieves the astounding task of comprehending and producing the meanings of human language in all their glory.

REFERENCES


