Neuroscience and Biobehavioral Reviews 81 (2017) 194–204

Contents lists available at ScienceDirect

Neuroscience and Biobehavioral Reviews

journal homepage: www.elsevier.com/locate/neubiorev

Review article

The core and beyond in the language-ready brain

Peter Hagoort

Max Planck Institute for Psycholinguistics; Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands

ARTICLE INFO

Article history:
Received 23 August 2016
Received in revised form 18 January 2017
Accepted 20 January 2017
Available online 11 February 2017

Keywords:
Neurobiology of language
Memory
Unification
Syntax
Semantics
Pragmatics

ABSTRACT

In this paper a general cognitive architecture of spoken language processing is specified. This is followed by an account of how this cognitive architecture is instantiated in the human brain. Both the spatial aspects of the networks for language are discussed, as well as the temporal dynamics and the underlying neurophysiology. A distinction is proposed between networks for coding/decoding linguistic information and additional networks for getting from coded meaning to speaker meaning, i.e. for making the inferences that enable the listener to understand the intentions of the speaker.

Contents

1. Introduction

In this paper I will sketch in very general terms the cognitive architecture of both language comprehension and production, as well as the neurobiological infrastructure that makes the human brain language-ready. The focus will be on spoken language, since that is the language modality for which the brain has been adapted. However, it is worth bearing in mind that humans can also interface with language as a cognitive system using sign and text (visual); that is to say, the system can connect with input/output processes in any sensory modality. Language processing consists of a complex set of procedures to get from sound to meaning (in comprehension) or meaning to sound (in production), with remarkable speed and accuracy. In the first section, I outline in brief a selection of the major constituent operations during (de)coding propositional content. In the next section, I turn to the neurobiological infrastructure hypothesized to form the basis for language processing. The principal insights are inspired by the notion of "brain networks" for syntactic processing and the construction of meaning. In the final sections, I discuss the highly dynamic nature of language processing in relation to underlying neurophysiology, and,

E-mail address: peter.hagoort@donders.ru.nl

http://dx.doi.org/10.1016/j.neubiorev.2017.01.048
0149-7634/

importantly, the necessity of the involvement of brain areas beyond the core networks for coding/decoding with the aim to understand the speaker's message.

2. The cognitive architecture of language comprehension and production

2.1. The comprehension of spoken language

The first requirement when listening to speech is that the continuous speech input is perceptually segmented into discrete entities (features, segments, syllables) that can be mapped onto, and will activate, abstract phonological representations that are stored in long-term memory. It has been known since long (cf. the Cohort Model: Marslen-Wilson, 1984; TRACE: McClelland and Elman, 1986; Shortlist: Norris, 1994) that during word recognition the incoming and unfolding acoustic input (e.g., the word-initial segment *ca*. . ..) activates in parallel not only one but a whole set of lexical candidates (e.g., *captain, capture, captivate, capricious*. . .). This set of candidates is reduced, based on further incoming acoustic input and contextually based predictions, to the one that fits best. This word recognition process happens extremely fast, and is completed within a few hundred milliseconds, whereby the exact duration is co-determined by the moment in time at which a particular word form deviates from all others in the mental lexicon of the listener (the so-called recognition point). Given the rate of typical speech (~4–6 syllables per second), one can deduce that word recognition is extremely fast and efficient, taking no more than 200–300 ms.

Importantly, achieving the mapping from acoustics to stored abstract representation is not the only aspect of lexical processing. For example, words are not processed as unstructured entities. Based on the morpho-phonological characteristics of a given word, a process of lexical decomposition takes place in which stems and affixes are separated. For spoken words, the trigger for decomposition can be something as simple as the inflectional rhyme pattern (IRP), which is a phonological pattern signaling the potential presence of an affix (Bozic et al., 2010). Interestingly, words seem to be decomposed by rule; that is to say, the decompositional, analytic processes are triggered for words with obvious parts (e.g., teacup = tea-cup; uninteresting = un-inter-est-ing) but also for semantically opaque words (e.g., bell-hop), and even nonwords with putative parts (e.g., blicket-s, blicket-ed). Decomposing lexical input appears to be a ubiquitous and mandatory perceptual strategy (e.g. Fiorentino and Poeppel, 2007; Solomyak and Marantz, 2010). Many relevant studies, especially with a view toward neurocognitive models, are reviewed by Marslen-Wilson (2007).

Recognizing word forms is an entrance point for the retrieval of syntactic (lemma) and semantic (conceptual) information. Here, too, the process is cascaded in nature. That is, based on partial phonological input, the meanings of multiple lexical candidates are co-activated (Zwitserlood, 1989). Multiple activation has so far not been found for lemma information that specifies the syntactic features (e.g., word class, grammatical gender) of a lexical entry. The reason might be that the phrase structure context provides strong constraints for the syntactic slot (say, a noun or a verb) that will be filled by the current lexical item (Lau et al., 2006; Müller and Hagoort, 2006). To what degree lemma and concept retrieval are sequential or parallel in nature during on-line comprehension, is not clear either. Results from electrophysiological recordings (ERPs), however, indicate that most of the retrieval and integration processes are completed within 500 ms (Kutas and Federmeier, 2011; see also below).

The processes discussed so far all relate to the retrieval of information from what is referred to as the mental lexicon. This is the information that in the course of language acquisition gets encoded and consolidated in neocortical memory structures, mainly located in the temporal and part of the parietal lobes. However, language processing is (i) more than memory retrieval and (ii) more than the simple concatenation of retrieved lexical items. The expressive power of human language (its generative capacity) derives from being able to combine elements from memory in endless, often novel ways. This process of deriving complex meaning from lexical building blocks will be referred to as unification (Hagoort, 2005). As we will see later, (left) frontal cortex structures are implicated in unification.

In short, the cognitive architecture necessary to realize the expressive power of language is tripartite in nature, with levels of form (speech sounds, or graphemes in text, or manual gestures in sign language), syntactic structure, and meaning as the core components of our language faculty (Chomsky, 1965; Jackendoff, 1999; Levelt, 1989, 1999). These three levels are domain-specific but at the same time interacting during incremental language processing. The principle of compositionality is often invoked to characterize the expressive power of language at the level of meaning. A strict account of compositionality states that the meaning of an expression is a function of the meanings of its parts and the way they are syntactically combined (Fodor and Lepore, 2002; Heim and Kratzer, 1998; Partee, 1984). In this account, complex meanings are assembled bottom-up from the meanings of the lexical building blocks via the combinatorial machinery of syntax. This is sometimes referred to as simple composition (Jackendoff, 1997). That some operations of this type are required is illustrated by the obvious fact that the same lexical items can be combined to yield different meanings: *dog bites man* is not the same as *man bites dog*. Syntax matters. It matters, however, not for its own sake but in the interest of mapping grammatical roles (subject, object) onto thematic roles (agent, patient) in comprehension, and in the reverse order in production. The thematic roles will fill the slots in the situation model/event schema (specifying states and events) representing the intended message.

However, that this account is not sufficient can be seen in adjective-noun expressions such as "flat tire" "flat beer," "flat note," etc. (Keenan, 1979). In all these cases, the meaning of "flat" is quite different and strongly context dependent. Thus, structural information alone will need to be supplemented. On its own it does not suffice for constructing complex meaning on the basis of lexical-semantic building blocks. Moreover, ERP (and behavioral) studies have found that non-linguistic information that accompanies the speech signal (such as information about the visual environment, about the speaker, or about co-speech gestures; Van Berkum et al., 2008; Willems et al., 2007, 2008) are unified in parallel with linguistic sources of information. Linguistic and non-linguistic information conspire to determine the interpretation of an utterance on the fly. This all happens extremely fast, usually in less than half a second. Ultimately, formal accounts of unification, therefore, need to handle multiple, parallel streams of information.

We have made a distinction between memory retrieval and unification operations. Here I will sketch in more detail the nature of unification in interaction with memory retrieval. Based on the Chomskyan tradition in linguistics, processing models of unification have often focused on syntactic analysis. However, as we saw above, unification operations take place not only at the syntactic processing level. Combinatoriality is a hallmark of language across representational domains (cf. Jackendoff, 2002). Thus, at the semantic and phonological levels, too, lexical elements need to be combined and integrated into larger structures (cf. Hagoort, 2005). Nevertheless, models of unification are more explicit for syntactic processing than for semantic processing. For the syntactic level of analysis, we can therefore illustrate the distinction between memory retrieval and unification most clearly. According to the model

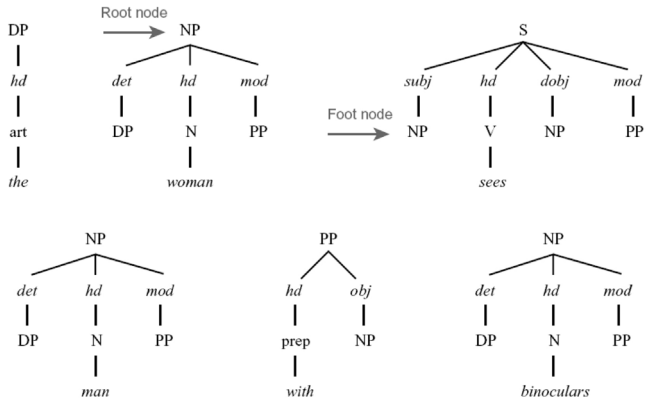

Fig. 1. Syntactic frames in memory. Frames such as these are retrieved on the basis of incoming word form information (the, woman, etc). DP: Determiner Phrase; NP: Noun Phrase; S: Sentence; PP: Prepositional Phrase; art: article; hd: head; det: determiner; mod: modifier; subj: subject; dobj: direct object. The head of a phrase determines the syntactic type of the frame (e.g. Noun for a Noun Phrase, Preposition for a Prepositional Phrase).

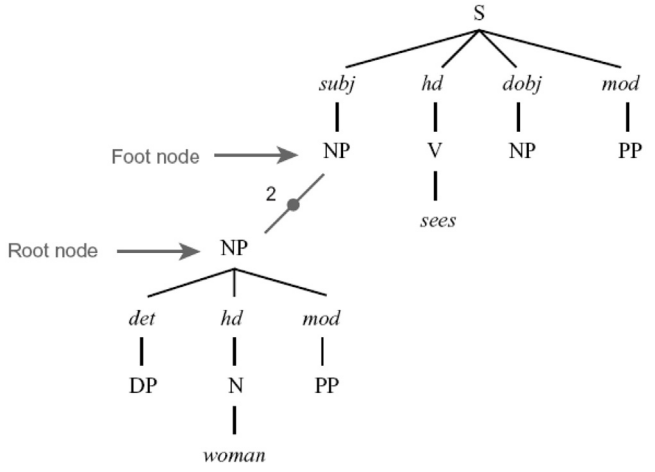

Fig. 2. The unification operation of two lexically specified syntactic frames. The unification takes place by linking the root node NP to an available foot node of the same category. The number 2 indicates that this is the second link that is formed during on-line processing of the sentence "The woman sees the man with the binoculars".

that I advocate (Hagoort, 2005, 2013), each word form in the mental lexicon is associated with a structural frame (Joshi and Schabes, 1997; Vosse and Kempen, 2000). This structural frame consists of a three-tiered unordered tree, specifying the possible structural environment of the particular lexical item (see Fig. 1). The top layer of the frame consists of a single phrasal node (e.g., NP). This so-called root node is connected to one or more functional nodes (e.g., Subject, Head, Direct Object) in the second layer of the frame. The third layer contains again phrasal nodes to which lexical items or other frames can be attached.

This parsing account is "lexicalist" in the sense that all syntactic nodes (e.g., S, NP, VP, N, V) are retrieved from the mental lexicon. In other words, chunks of syntactic structure are stored in memory. There are no syntactic rules that introduce additional nodes, such as in classical rewrite rules in linguistics (S –> NP VP). In the on-line comprehension process, structural frames associated with the individual word forms incrementally enter the unification workspace. In this workspace, constituent structures spanning the whole utterance are formed by a unification operation (see Fig. 2). This operation consists of linking up lexical frames with identical root and foot nodes, and checking agreement features (number, gender, person, etc.). Although the lexical-syntactic frames might differ between languages, as well as the ordering of the trees, what is claimed to be universal is the combination of lexically specified syntactic templates and unification procedures. Moreover across languages the same distribution of labour is predicted between brain areas involved in memory and brain areas that are crucial for unification.

The unification links between lexical frames are formed dynamically with initially competing activations for alternative linking possibilities. By consequence the strength of the unification links varies over time until a state of equilibrium is reached. Due to the inherent ambiguity in natural language, alternative unification candidates will usually be available at any point in the parsing process. That is, a particular root node (e.g., PP) often finds more than one matching foot node (i.e. PP) (see Fig. 2) with which it can form a unification link (Hagoort, 2003).

Ultimately, at least for sentences that do not tax the processing resources very strongly, one phrasal configuration results. This requires that among the alternative binding candidates only one remains active. The required state of equilibrium is reached through a process of lateral inhibition between two or more alternative unification links. In general, due to gradual decay of activation more recent foot nodes will have a higher level of activation than the ones that entered the unification space earlier. In addition, strength levels of the unification links can vary as a function of plausibility (semantic) effects. For instance, if instrumental modifiers under S-nodes have a slightly higher default activation than instrumental modifiers under an NP-node, lateral inhibition can result in overriding a recency effect.

The picture that is sketched above is based on the assumption that we always create a fully unified structure. This is, however, unlikely. In actual language processing oftentimes the comprehension system will work with bits and pieces (e.g. syntactic frames) that are not all unified into one fully specified phrasal configuration. Given both extralinguistic and language-internal contextual constraints and redundancy this is in the majority of cases still good enough to derive the intended message (see below).

The Unification Model as formalized in Vosse and Kempen (2000) has nevertheless a certain psychological plausibility. It accounts for sentence complexity effects known from behavioral measures, such as reading times. In general, sentences are harder to analyze syntactically when more potential unification links of similar strength enter into competition with each other. Sentences are easy when the number of U-links is small and of unequal strength. In addition, the model accounts for a number of other experimental findings in psycholinguistic research on sentence processing, including syntactic ambiguity (attachment preferences; frequency differences between attachment alternatives), and lexical ambiguity effects. Moreover, it accounts for breakdown patterns in agrammatic sentence analysis (for details, see Vosse and Kempen, 2000).

In my specification of the processing steps I have implicitly assumed that ultimately decoding the meaning is what language comprehension is about. However, while this might be a necessary aspect, it cannot be the whole story. Communication goes further than the exchange of explicit propositions. In essence it is a way to either change the mind of the listener, or to commit the addressee to the execution of certain actions, such as closing the window in reply to the statement "It is cold here". In other words, a theory of speech acts is required to understand how we get from coded meaning to inferred speaker meaning (cf. Grice, 1989).

Another assumption that I made, but which might also be incorrect as we saw above, relates to how much of the input the listener/reader analyzes. In classical models of sentence comprehension – of either the syntactic-structure-driven variety (Frazier, 1987) or in a constraint-based framework (Tanenhaus et al., 1995) – the implicit assumption is usually that a full phrasal configuration

results and a complete interpretation of the input string is achieved. However, oftentimes the listener arrives at an interpretation that is only partially consistent with the input string. This could be taken to indicate that the listener is interpreting the input on the basis of bits and pieces that are only partially analyzed. As a consequence, the listener might overhear semantic information (cf. the Moses illusion; Erickson and Mattson, 1981) or syntactic information (cf. the Chomsky illusion; Wang et al., 2011). In the question "How many animals of each kind did Moses take on the ark?, people often answer 'two', without noticing that it was Noah who was the guy with an ark, and not Moses. I found that likewise syntactic violations might go unnoticed if they are in a sentence constituent that provides no new information (Wang et al., 2011). Ferreira et al. (2002) introduced the phrase "good-enough processing" to refer to the listeners' and readers' interpretation strategies. In a good-enough processing context, linguistic devices that highlight the most relevant parts of the input might help the listener/reader in allocating processing resources optimally. This aspect of linguistic meaning is known as "information structure" (Büring, 2007; Halliday, 1967; Seuren, 1985). The information structure of an utterance essentially focuses the listener's attention on the crucial (new) information in it. In languages such as English and Dutch, prosody plays a crucial role in marking information structure. For instance, in question-answer pairs, the new or relevant information in the answer will typically be pitch accented. After a question like *What did Mary buy at the market?*, the answer might be *Mary bought VEGETABLES* (accented word in capitals). In this case, the word "vegetables" is the focus constituent, which corresponds to the information provided for the Wh-element in the question. There is no linguistic universal for signaling information structure. The way information structure is expressed varies within and across languages. In some languages it may impose syntactic locations for the focus constituent, in other languages focus-marking particles are used, or prosodic features like phrasing and accentuation (Kotschi, 2006; Miller, 2006).

In summary, language comprehension requires an analysis of the input that allows the retrieval of relevant information from memory (the mental lexicon). The lexical building blocks are unified into larger structures decoding the propositional content. Further inferential steps are required to derive the intended message of the speaker from the coded meaning. Based on the listener's comprehension goals, the input is analyzed to a lesser or larger extent. Linguistic marking of information structure co-determines the depth of processing of the linguistic input. In addition, non-linguistic input (e.g., co-speech gestures, visual context) is immediately integrated into the situation model that results from processing language in context.

2.2. Producing language

While one can describe speech comprehension as the mapping from sound to meaning, in speaking we travel the processing space in the reverse order. In speaking, a preverbal message is transformed by a series of computations into a linearized sequence of speech sounds (for details, see Levelt, 1989, 1999). As in language comprehension this requires the retrieval of building blocks from memory and their unification at multiple levels. Most research on speaking has focused on single word production, using experimental paradigms such as picture naming. The whole cascade of word production from the stage of conceptual preparation to the final articulation happens in about 600 ms (Indefrey and Levelt, 2004). Since we perform this process in an incremental fashion, we can easily speak 2–4 words per second. Moreover, this is done with amazing efficiency, since on average only once in a thousand words do we make a speech error (Bock, 2011; Deese, 1984). The whole cascade of processes starts with a preverbal message that triggers the selection of the required lexical concepts, i.e., the concepts for which a word form is available in the mental lexicon. The activation of a lexical concept leads to the selection of the target lemma, which gets phonologically encoded. At the stage of lemma selection, morphological unification of, for instance, stem and affix takes place. Once the phonological word forms are retrieved, they will result in the retrieval and unification of the syllables that compose a phonological word in its current speech context.

Next to the selection of lexical concepts, structures have to be encoded in which the lexical concepts can be included. This requires the availability of abstract structures for shaping sentence form or, alternatively, lexical-syntactic building blocks (Vosse and Kempen, 2000) that guide the assembly of larger structures in production (cf. Konopka and Meyer, 2014).

Although speech comprehension and speaking recruit many of the same brain areas during sentence-level semantic processes, syntactic operations and lexical retrieval (Menenti et al., 2011), there are still important differences. The most important difference is that although speakers pause, repair, etc., they nevertheless cannot bypass syntactic and phonological encoding of the utterance that they intend to produce. What is good enough for the listener is often not good enough for the speaker, and although ambiguity is a key feature of language comprehension at multiple levels, it plays only a minor role in production. Nevertheless production and comprehension operate in an interdependent way. For instance, it might well be that the interconnectedness of the cognitive and neural architectures for language comprehension and production enables the production system to participate in generating internal predictions while in the business of comprehending linguistic input (Dell and Chang, 2014).

3. The neurobiological infrastructure

Classically, and mainly based on evidence from deficits in aphasic patients, the perisylvian cortex in the left hemisphere has been seen as the crucial network for supporting the processing of language. For a long time, the key components were assumed to be Broca's area in the left inferior frontal cortex and Wernicke's area in the left superior temporal cortex, with these areas mutually connected by the arcuate fasciculus. These areas and their roles in language comprehension and production are often still described as the core language nodes in handbooks on brain function (see Fig. 3). However, in later times patient studies and especially recent neuroimaging studies in healthy subjects have revealed that (i) the distribution of labour between Broca's and Wernicke's area is different than proposed in the classical model, and (ii) a much more extended network of areas is involved, not only in the left hemisphere, but also involving homologous areas in the right hemisphere. One alternative account to the classical view is the Memory, Unification and Control (MUC) model proposed by Hagoort (2005, 2013). In this model, the distribution of labour is as follows (see Fig. 4): Areas in the temporal cortex (in yellow) and parts of parietal cortex subserve the knowledge representations that during acquisition have been laid down in memory. These areas store information about word form, word meanings and the syntactic templates that were discussed above. Dependent on information type, different parts of temporal/parietal cortex are involved. Frontal cortex areas (Broca's area and adjacent cortex; in blue) are crucial for the unification operations. These operations generate larger structures from the building blocks that are retrieved from memory. In addition, executive control is involved, for instance when the correct target language is selected or turn taking in conversation is orchestrated. Control areas involve dorsolateral prefrontal cortex (in pink) and a midline structure known as the Anterior Cingulate Cortex (not shown in Fig. 4).

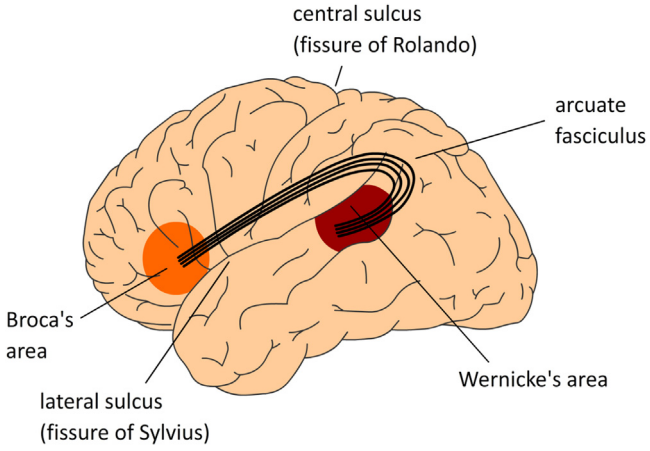

Fig. 3. The classical Wernicke-Lichtheim-Geschwind model of the neurobiology of language. In this model Broca's area is crucial for language production, Wernicke's area subserves language comprehension, and the necessary information exchange between these areas (such as in reading aloud) is done via the arcuate fasciculus, a major fibre bundle connecting the language areas in temporal cortex (Wernicke's area) and frontal cortex (Broca's area). The language areas are bordering one of the major fissures in the brain, the so-called Sylvian fissure. Collectively, this part of the brain is often referred to as perisylvian cortex. Reprinted with permission from Hagoort (2014).

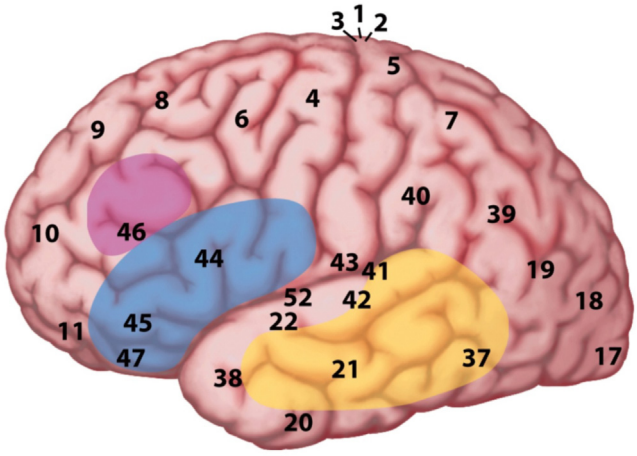

Fig. 4. The MUC model of language. The figure displays a lateral view of the left hemisphere. The numbers indicate Brodmann areas. These are areas with differences in the cytoarchitectonics (i.e. composition of cell types). The memory areas are in the temporal cortex (in yellow). Unification requires the contribution of Broca's area (Brodmann areas 44 and 45) and adjacent cortex (Brodmann areas 47 and 6) in the frontal lobe. Control operations recruit another part of the frontal lobe (in pink), and the Anterior Cingulate Cortex (ACC; not shown in the figure). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

In the next sections I will discuss in more detail some of the brain networks supporting different types of information that are crucial for language. For the three core types of information (phonological, syntactic, semantic), the same general distinction can be made between retrieval operations and unification. Retrieval refers to accessing language-specific information in memory. Unification is the (de)composition of larger structures from the building blocks that are retrieved from memory. In more computational terms, memory and unification refer to the distinction between look-up tables and compact procedures (Gallistel and King, 2010).

A *meta*-analysis of a large number of neuroimaging studies (Hagoort and Indefrey, 2014) reveal a gradient in left-temporofrontal cortex for semantic and syntactic processing operations. Higher syntactic demands most reliably activate the more dorsal parts of posterior LIFG (BA 44/45), the right posterior IFG and the left posterior STG and MTG. In addition there is reliable activation of the left precuneus, the left inferior parietal lobule, and the right posterior MTG. Higher semantic demands most reliably activate all parts of posterior LIFG (but BA 45/47 are reported twice as often as BA 44), the right posterior IFG and the left middle and posterior MTG. In addition, there is a reliable activation of the medial prefrontal cortex that is not activated in syntactic processing, as well as activations of the left anterior insula, angular gyrus and the posterior ITG.

3.1. The syntactic network

In comparison with phonological and semantic processing, which have compelling bilateral contributions (in contrast to the classical left-hemisphere-only model), syntactic processing seems strongly lateralized to the left hemisphere perisylvian regions. Independent of the input modality (visual in reading, auditory in speech). two supramodal areas for syntactic processing are the left posterior superior/middle temporal gyrus (STG/MTG) and the left inferior frontal cortex. The left posterior temporal cortex is known to be involved in lexical processing (Hickok and Poeppel, 2004, 2007; Indefrey and Cutler, 2004; Lau et al., 2006). In connection to the Unification Model, this part of the brain might be important for the retrieval of the syntactic frames that are stored in the lexicon. The Unification Space, where individual frames are connected into a phrasal configuration for the whole utterance, might recruit the contribution of Broca's area (left inferior frontal cortex, LIFC). As we have seen above in the formal account of syntactic unification, competition and selection are an integral part of the unification operation. Hence this view is not inconsistent with the idea that the LIFC plays an important role in selection (Thompson-Schill et al., 2005). However, the MUC model stresses the idea that in processing language selection operates in the service of unification.

Direct empirical support for the proposed distribution of labor between LIFC (Broca's area) and temporal cortex was found in a study of Snijders et al. (2009). In their fMRI study participants read sentences and word sequences containing word-category (noun-verb) ambiguous words at critical position (e.g., "watch"). Regions contributing to the syntactic unification process should show enhanced activation for sentences compared with words, and only within sentences display a larger signal for ambiguous than unambiguous conditions. The posterior LIFC showed exactly this predicted pattern, confirming the hypothesis that LIFC contributes to syntactic unification. The left posterior middle temporal gyrus was activated more for ambiguous than unambiguous conditions, as predicted for regions subserving the retrieval of lexical-syntactic information from memory. It thus seems that the left inferior frontal cortex is crucial for syntactic processing in conjunction with the left posterior middle temporal gyrus (Goucha and Friederici, 2015; Friederici, this issue), a finding supported by patient studies with lesions in these very same areas (Caplan and Waters, 1996; Rodd et al., 2010; Tyler et al., 2011).

3.2. The semantic network

In recent years, there has been growing interest in investigating the cognitive neuroscience of semantic processing. A series of fMRI studies aimed at identifying the semantic processing network. These studies either compared sentences containing semantic/pragmatic anomalies with their correct counterparts (e.g., Friederici et al., 2003; Hagoort et al., 2004; Kiehl et al., 2002; Ruschemeyer et al., 2006) or compared sentences with and without semantic ambiguities (Davis et al., 2007; Hoenig and Scheef, 2005; Rodd et al., 2005). The most consistent finding across all of these studies is the activation of the left inferior frontal cortex (LIFC),

more in particular BA 47 and BA 45. In addition, the left superior and middle temporal cortices are often found to be activated, as well as left inferior parietal cortex. For instance, Rodd and colleagues found that sentences with lexical ambiguities resulted in increased activations in LIFC and in the left posterior middle/inferior temporal gyrus. In this experiment all materials were well-formed sentences in which the ambiguity usually goes unnoticed. Nevertheless, very similar results were obtained as in experiments that used semantic anomalies. Areas involved in semantic unification were found to be sensitive to the increase in semantic unification load due to the ambiguous words. Semantic unification can be seen as filling the slots in an abstract event schema, where in the case of multiple word meanings for a given lexical item competition and selection increase in relation to filling a particular slot in the event schema. As with syntactic unification, the availability of multiple candidates for a slot will increase the unification load. In the case of the lexical ambiguities there is no syntactic competition, since both readings activate the same syntactic template (in this case the NP-template). Increased processing is hence due to integration of meaning instead of syntax.

In short, the semantic processing network seems to include at least LIFC, left superior/middle temporal cortex, and the (left) inferior parietal cortex. To some degree, the right hemisphere homologues of these areas are also found to be activated. Below we will discuss the possible contributions of these regions to semantic processing.

An indication for the respective functional roles of the left frontal and temporo-parietal cortices in semantic unification comes from a few studies investigating semantic unification of multimodal information with language. Using fMRI, Willems and colleagues assessed the neural integration of semantic information from spoken words and from co-speech gestures into a preceding sentence context (Willems et al., 2007). Spoken sentences were presented in which a critical word was accompanied by a co-speech gesture. Either the word or the gesture could be semantically incongruous with respect to the previous sentence context. Both an incongruous word as well as an incongruous gesture led to increased activation in LIFC as compared to congruous words and gestures (for a similar finding with pictures of objects, see Willems et al., 2008). Interestingly, the activation of the left posterior superior temporal cortex was increased by an incongruous spoken word, but not by an incongruous hand gesture (Willems et al., 2007). This suggests that activation increases in left posterior temporal cortex are triggered most strongly by processes involving the retrieval of lexical-semantic information. LIFC, on the other hand, is a key node in the semantic unification network, unifying semantic information from different modalities. From these findings it seems that semantic unification is realized in a dynamic interplay between LIFC as a multimodal unification site on the one hand, and modality specific areas on the other hand.

Importantly, semantic processing is more than the concatenation of lexical meanings. Over and above the retrieval of individual word meanings, sentence and discourse processing requires combinatorial operations that result in a coherent interpretation of multi-word utterances. These operations do not adhere to principle of strict compositionality. World knowledge, information about the speaker, co-occurring visual input and discourse information all trigger similar neural responses as sentence-internal semantic information. A network of brain areas, including the left inferior frontal cortex, the left superior/middle/inferior temporal cortex, the left inferior parietal cortex and, to a lesser extent, their right hemisphere homologues are recruited to perform semantic unification. The general finding is that semantic unification operations are under top-down control of left, and in the case of discourse, also right inferior frontal cortex. This contribution modulates activations of lexical information in memory as represented by the left superior and middle temporal cortex, with presumably additional support for unification operations in left inferior parietal areas (e.g., angular gyrus). These findings also imply that the slots in an event schema can be filled with information from different input formats. A more formal account of such a multimodal semantic unification process is, however, still lacking.

4. The network topology of the language-ready brain

We have seen that the language network in the brain is much more extended than was thought for a long time, and not only includes areas in the left hemisphere but also right hemisphere areas. However, the evidence of additional activations in the right hemisphere and areas other than Broca's and Wernicke's area, does not take away the bias in favor of left perisylvian cortex. In a meta-analysis based on 128 neuroimaging papers, Vigneau et al. (2010) compared left and right hemisphere activations related to language processing. On the whole, for phonological processing, lexico-semantic processing and sentence or text processing, the activation peaks in the right hemisphere comprised less than one third of the activation peaks in the left hemisphere. Moreover in the large majority of cases the right hemisphere activations were in homotopic areas, suggesting strong inter-hemispheric influence. It is therefore justified to think that for the large majority of the population (e.g., with the exception of some portion of left-handers), the language readiness of the human brain is to a large extent dependent on the organization of the left perisylvian cortex. Especially, the network organization of the left perisylvian cortex has been found to show characteristics that distinguishes it from the right perisylvian cortex – and from homologue areas in other primates. A recent technique for tracing fiber bundles in the living brain is Diffusion Tensor Imaging (DTI). Using DTI, Rilling et al. (2008) tracked the arcuate fasciculus in humans, chimpanzees and macaques. These authors found in humans a prominent temporal lobe projection of the arcuate fasciculus that is much smaller or absent in nonhuman primates. Moreover, connectivity with the middle temporal gyrus (MTG) was more widespread and of higher probability in the left than in the right hemisphere. This human specialization may be relevant for the evolution of language. Catani et al. (2007) found that the human arcuate fasciculus is strongly lateralized to the left, with quite some variation on the right. On the right, some people lack an arcuate fasciculus, in others it is smaller in size, and only in a minority of the population this fiber bundle is of equal size in both hemispheres. This pattern of lateralization was confirmed in a study on 183 healthy right-handed volunteers in the age range between 5 and 30 years (Lebel and Beaulieu, 2009). In this study the lateralization pattern did not differ with age or gender. The arcuate fasciculus lateralization is present at 5 years and remains constant throughout adolescence into adulthood. However, another recent study comparing 7 year olds with adults (Brauer et al., 2011) reports that the arcuate fasciculus is still relatively immature in the children compared with the adults.

In addition to the arcuate fasciculus, other fiber bundles are important as well in connecting frontal with temporoparietal language areas (see Fig. 5). These include the superior longitudinal fasciculus (adjacent to the arcuate fasciculus) and the extreme capsule fasciculus as well as the uncinate fasciculus, connecting Broca's area with superior and middle temporal cortex along a ventral path (Anwander et al., 2007; Friederici, 2009; Kelly et al., 2010).

DTI is not the only way to trace brain connectivity. It has been found that imaging the brain during rest reveals low-frequency (<0.1 Hz) fluctuations in the fMRI signal. These fluctuations are correlated across areas that are functionally related (Biswal et al., 1995; Biswal and Kannurpatti, 2009). This so-called resting state fMRI can thus be used as an index of functional connectivity.

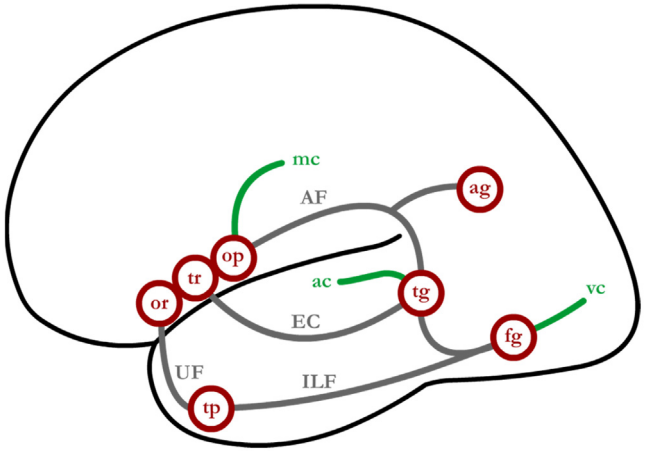

Fig. 5. Simplified illustration of the anatomy and connectivity of the left hemisphere language network. Cortical areas are represented as red circles: pars orbitalis (or), pars triangularis (tr) and pars opercularis (op) of the left inferior frontal cortex (LIFC); angular gyrus (ag), superior and middle temporal gyri (tg), fusiform gyrus (fg) and temporal pole. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

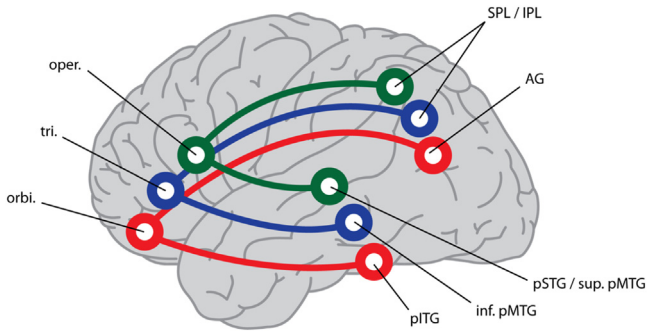

Fig. 6. A schematic drawing of the topographical connectivity pattern between frontal and temporal/parietal cortex in the perisylvian language network, as revealed by resting state fMRI (after Xiang et al., 2010). The strongest connections to the pars opercularis (oper.), part triangularis (tri.) and pars orbitalis (orbi.) of Broca's region are shown. SPL/IPL: Superior Parietal Lobule/Inferior Parietal Lobule; AG: Angular Gyrus; pSTG: posterior Superior Temporal Gyrus; sup. pMTG: superior posterior Middle Temporal Gyrus; inf. pMTG: inferior posterior Middle Temporal Gyrus; pITG: posterior Inferior Temporal Gyrus. Reprinted with permission from Hagoort (2014).

Although both DTI and resting state fMRI measure connectivity, in the case of DTI the connectivity can often be related to anatomically identifiable fibre bundles. Resting state connectivity measures the functional correlations between areas without providing a correlate in terms of an anatomical tract. Using the resting state method, Xiang et al. (2010) found a clear topographical functional connectivity pattern in the left inferior frontal, parietal, and temporal areas (see Fig. 6). In the left – but not the right – perisylvian cortex, functional connectivity patterns obeyed the tripartite nature of language processing (phonology, syntax and semantics). These results support the assumption of the functional division for phonology, syntax, and semantics of the left inferior frontal cortex, including Broca's area, as proposed by the MUC model (Hagoort, 2005), and revealed a topographical functional organization in the left perisylvian language network, in which areas are most strongly connected according to information type (i.e., phonological, syntactic, and semantic).

In summary, despite increasing evidence of right hemisphere involvement in language processing, it still seems clear that the left perisylvian cortex has certain network features that stand out in comparison to other species, and make it especially suited for supporting the tripartite architecture of human language.

5. Neurophysiology and timing

Although I have thus far emphasized functional neuroanatomy and the insights from imaging, it is worth bearing in mind what electrophysiological data add to the functional interpretations we must entertain. As I discussed at the outset, one of the most remarkable characteristics of speaking and listening is the speed at which it occurs. Speakers produce easily between 2 and 5 words per second; information that has to be decoded by the listener within roughly the same time frame. Considering that the acoustic duration of many words is in the order of a few hundred milliseconds, the immediacy of the electrophysiological language-related effects is remarkable. For instance, the early left anterior negativity (ELAN), a syntax-related effect (Friederici et al., 2003) has an onset on the order 100–150 ms after the acoustic word onset. The onset of the N400 is approximately at 250 ms, and another language relevant ERP, the so-called, P600, usually starts at about 500 ms. The majority of these effects thus happen well before the end of a spoken word. Classifying visual input (e.g., a picture) as depicting an animate or inanimate entity takes the brain approximately 150 ms (Thorpe et al., 1996). Roughly the same amount of time is needed to classify orthographic input as a letter (Grainger et al., 2008). If one takes this as the reference time, the early appearance of an ELAN response to a spoken word is remarkable. In physiological terms, it might be just too fast for long-range recurrent feedback to have its effect on parts of primary and secondary auditory cortex involved in first-pass acoustic and phonological analysis. Recent modeling work suggests that early ERP effects are best explained by a model with feedforward connections only. Backward connections become essential only after 220 ms (Garrido et al., 2007). The effects of backward connections are, therefore, not manifest in the latency range of at least the ELAN, since not enough time has passed for return activity from higher levels. But also the N400 follows the word recognition points closely in time in the case of speech. This suggests that what is going on in on-line language comprehension is presumably for a substantial part based on predictive processing. Under most circumstances, there is simply not enough time for top-down feedback to exert control over an ongoing bottom-up analysis. Very likely, lexical, semantic and syntactic cues conspire to predict characteristics of the next anticipated word, including its syntactic and semantic make-up. A mismatch between contextual prediction and the output of bottom-up analysis results in an immediate brain response recruiting additional processing resources for the sake of salvaging the on-line interpretation process. Recent ERP studies have provided evidence that context can indeed result in predictions about a next word's syntactic features (i.e., gender; Van Berkum et al., 2005) and word form (DeLong et al., 2005). Lau et al. (2006) provided evidence that the ELAN elicited by a word category violation was modulated by the strength of the expectation for a particular word category in the relevant syntactic slot. In summary, predictive coding is likely a central feature of the neurocognitive infrastructure for language.

5.1. The dynamic interplay between memory and unification

Although I have made a connection between functional components of the cognitive architecture for language and specific brain regions, this is an idealization of the real neurophysiological dynamics of the perisylvian language network. Crucially, for language as for most other cognitive functions, the functional contribution of any area or region has to be characterized in the context of the network as a whole, where specialization of any node is

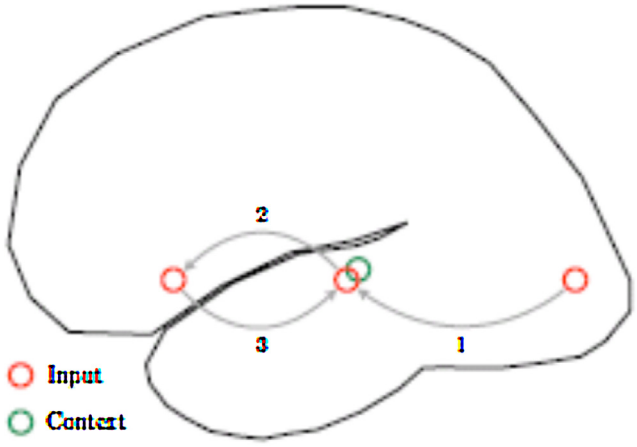

Fig. 7. Processing cycle subserving word meaning comprehension in the left hemisphere language network. Inputs are conveyed from sensory regions (here visual cortex) to the inferior, middle and superior temporal gyri (1), where lexical information is activated. Signals are hence relayed to the inferior frontal gyrus (2), where neurons respond with a sustained firing pattern. Signals are then fed back into the same regions in temporal cortex from where they were received (3). A recurrent network is thus set up, which allows information to be maintained on-line, a context (green circle) to be formed during subsequent processing cycles, and incoming words to be unified within the context. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

only relative and realized in a dynamic interaction with the other nodes in the network (Mesulam, 1990, 1998). This will be illustrated below on the basis of a neurophysiological account of the N400, the most well established ERP effect related to language (Kutas and Hillyard, 1980), and more in particular to semantic unification. Similar accounts are to be made for syntactic and phonological unification.

As I discussed above, in posterior and inferior temporal and parietal (angular gyrus) regions neuronal populations are activated that represent lexical information associated with the incoming word, including its semantic features. From here, neural signals can follow two routes. The first exploits local connectivity within these posterior regions, resulting in a graded activation of neighboring neuronal populations, coding for related lexical-semantic information. Such local spread of activation contributes to setting up a lexical-semantic context in temporo-parietal cortex (Fig. 7, green circle), and may underlie priming and pre-activation at short SOAs (Lau et al., 2008). The second route is based on long-distance connections to LIFC, through direct white matter fibers resulting in the selective activation of populations of frontal cortex neurons. These will respond with a self-sustaining firing pattern (see Durstewitz et al., 2000; for a review). Efferent signals in this case can only take the long-range route back. The most parsimonious option is that frontal neurons will send efferent signals back to the same regions in temporo-parietal cortex from where afferent signals were received. This produces another spread of activation to neighboring temporo-parietal regions, which implies that connections representing a given local semantic context will be strengthened. This may be related to priming and prediction at longer SOAs, when the contribution of LIFC is also more prominent (Lau et al., 2008). During each word processing cycle the memory (temporo-parietal) and unification (inferior frontal) components interact, by letting activation reverberate through the circuit in Fig. 7. Achieving the necessary outcomes for language comprehension may be more or less demanding, depending on how close the relation is between input and context, as we shall see below.

This description of a typical word processing cycle appears to be the simplest possible solution given constraints from brain imaging (the involvement of temporal, parietal, and inferior frontal regions), neuroanatomy (the existence of direct white matter pathways), and neurophysiology (persistent firing of LIFC neurons). However, the proposal

requires further elaboration, and a computational implementation that would confer a precise meaning to the envisaged processing steps.

Reverberation in the fronto-temporal circuit might be crucial for basic neurophysiological reasons. Friston (2005) assigns different roles to different neurotransmitters, depending on their decay times. Feedforward connections appear to mediate their post-synaptic effects through fast AMPA and $GABA_A$ receptors, and feedback connections are most probably mediated by much slower NMDA receptors. NMDA receptors are relatively frequent in supra-granular layers, where backward connections terminate (Bastos et al., 2012; Kiebel et al., 2008; Sherman and Guillery, 1998; Sherman, 2007). NMDA-mediated channels may have a role in relaying modulatory effects that are more extended in time (Wong and Wang, 2006). Lisman et al. (1998) have shown that NMDA-receptor mediated EPSPs are critical for the maintenance of information in working memory. They allow a network to maintain its active state without the need for synaptic modification. There is increasing evidence that cortical reverberation by re-entry is important for working memory (Fuster, 2009; Wang, 1999). Baggio and Hagoort (2011) hypothesize that the same is true for language. The feedforward pathways from temporal/parietal cortex to LIFC may be a rapid-decay route requiring NMDA mediated re-entry from LIFC to maintain lexical information active over time, as is essential for multi-word unification.

This neurophysiological account can serve as a basis for a neurocomputational model of the N400. In this proposal the N400 component reflects reverberating activity within the posterior-frontal network during one or perhaps several cycles, as shown in Fig. 7. Activity starts building up around 250 ms from word onset, reflecting the summation of post-synaptic currents injected by inferior temporal areas and by neighboring populations in MTG/STG. The direct white matter routes allow for a rapid spread of activation to LIFC. The peak of the N400 coincides with the completion of the cycle; that is with the re-injection of currents into temporal/parietal regions. Across several word-processing cycles, a pattern of neuronal activity emerges in these posterior regions, encoding a local context. This is the result of activation spreading to areas neighboring to those activated by the input during the feedforward sweep, and of a similar process taking place during the feedback from LIFC. This process strengthens learned associations between semantic features.

Consider now the case in which semantic relatedness is manipulated, as for instance in "The girl was writing a letter when her friend spilled coffee on the tablecloth/paper." (Baggio et al., 2008). Processing the fragment "The girl was writing a letter when her friend spilled coffee on the . . .̆sets up a context, maintained over time by input from LIFC. Semantic features associated with the words *writing* and *letter* are activated (Brunel and Lavigne, 2009; Cree et al., 1999; Cree and McRae, 2003; Masson, 1991; Masson 1995; McRae and Ross, 2004; Moss et al., 1994). If these include features that contribute to activating the concept of paper, then there will be some overlap between the neuronal populations representing the context and those that selectively respond to the given input, which is to the incoming word *paper*. Such overlap will be smaller for *tablecloth*. The larger the overlap is between context and input, the smaller the amplitude of the scalp-recorded ERP will be. In particular, the incoming word that benefits from a larger overlap with the context (*paper*) results in a smaller N400 compared to the word that leads to a smaller overlap (*tablecloth*). The inverse relation between semantic relatedness and N400 amplitude follows from an inverse relation between the degree of overlap of neuronal sources and the amplitude of scalp-recorded ERPs. The amplitude of any given

neuronal generator scales with the size of the contributing population of neurons that are concurrently activated. Under the assumption that there is an N400 unification effect, the increase in the N400 amplitude as a function of unification load can be explained as follows. Neuronal populations in LIFC (coding for the current non-local context), upon receiving input from temporal/parietal cortex, start firing in a sustained manner, and inject currents back into the same regions from where signals were received. In this way transient links are dynamically established between semantic types for which temporal and parietal cortex might be the hubs (convergence zones of distributed representations). This account is consistent with the finding that some of the strongest neuronal generators of N400 are localized in the left middle and superior temporal cortex. This is where most afferent signals are projected: (i) from peripheral areas via inferior temporal cortex during early processing stages (~200 ms); (ii) through local connectivity in MTG/STG due to spreading activation from input-selective populations to neighboring temporal areas; (iii) from LIFC during the feedback that supports unification and the on-line maintenance of context. LIFC may show a comparatively smaller net effect of post-synaptic currents over shorter time intervals, possibly due to fewer signals re-injected through local connectivity in LIFC itself, but a stronger activation (as revealed by metabolic measures) over longer time periods, due to the persistent firing patterns produced by LIFC neurons. This could explain why MEG/EEG source analyses may fail to reveal significant contributions of LIFC, whereas fMRI does show a strong response in LIFC. Also, the time-locking of neuronal responses appears to be sharper in posterior temporal cortex than in inferior frontal areas (Liljeström et al., 2009). Activity in LIFC is presumably relatively insensitive to the onset and offset times of the stimuli, and is rather a self-sustaining state which is relatively unaffected by trial-to-trial variation. In contrast, bottom-up activation in MTG/STG and adjacent regions may have tighter deadlines, partly due to the proximity to sensory areas.

This account of the N400 (for further details, see Baggio and Hagoort, 2011) is consistent with available anatomical and functional data, as well as the accounts proposed by Kutas and Federmeier in their review of thirty years N400 research (Kutas and Federmeier, 2011) and by Nieuwland et al. (2010). It explains the N400 as resulting from the summation of currents injected by frontal into temporal/parietal areas (unification) with currents that are already circulating within the latter regions due to the local spread of activation to neighboring neuronal populations. In real-time language processing access, selection, pre-activation and unification are all part of a word processing cycle; that is, a continuous pattern of neuronal activity unfolding over time within a distributed cortical network.

6. Beyond the core areas for language

In the previous sections I have focused on the neurobiological infrastructure for decoding/encoding propositional content. However, in many instances, linguistic expressions are underdetermined with respect to the meaning that they convey. What is said and what is understood are often not the same. Communication goes further than the exchange of explicit propositions. In essence the goal of the speaker is to either change the mind of the listener, or to commit the addressee to the execution of certain actions, such as closing the window in reply to the statement "It is cold here". In other words, a theory of speech acts is needed to understand how we get from coded meaning to inferred speaker meaning (cf. Grice, 1989; Hagoort and Levinson, 2015; Levinson, 2013). The steps that are required to get from coded meaning to speaker meaning involve additional neuronal networks, as recent studies have shown. In a study on conversational implicatures and indirect requests, for example, Bašnáková and coworkers (2014) contrasted direct and indirect replies – two classes of utterances whose speaker meanings are either very similar to, or markedly different from, their coded meaning. In their study participants listened to natural spoken dialogue in which the final and critical utterance, e.g., "It is hard to give a good presentation", had different meanings depending on the dialogue context and the immediately preceding question. This critical utterance either served as a direct reply (to the question "How hard is it to give a good presentation?"), or an indirect reply (to "Did you like my presentation?"). In the indirect reply condition, there were activations in the medial prefrontal cortex (mPFC), and in the right temporo-parietal junction (rTPJ), a pattern typical for tasks which involve making inferences about other minds (Amodio and Frith, 2006; Mitchell et al., 2006; Saxe et al., 2006). Although the exact role of all the individual theory-of-mind (ToM) regions is not yet clearly established, both mPFC and right TPJ constitute core regions in ToM tasks (Carrington and Bailey, 2009). The most specific hypothesis about the role of the posterior part of (right) TPJ (Mars et al., 2012) in the mentalizing network is that it is implicated in mental state reasoning, i.e. thinking about other people's beliefs, emotions and desires (Saxe, 2010). The mPFC activation in the Bašnáková et al. study (2014) was found in parts of the mPFC that are associated with complex socio-cognitive processes such as mentalizing or thinking about the intentions of others (such as communicative intentions, right anterior mPFC) or about oneself (right posterior mPFC). Interestingly, the involvement of these regions is also consistently observed in discourse comprehension (e.g. Mar, 2011; Mason and Just, 2009). This might come as no surprise, since it is likely that the motivations, goals, and desires of fictional characters are accessed in a similar manner as with real-life protagonists (Mar and Oatley, 2008). In fact, an influential model from the discourse processing literature (Mason and Just, 2009) ascribes the dorsomedial part of the frontal cortex and the right TPJ a functional role as a *protagonist perspective network*, which generates expectations about how the protagonists of stories will act based on understanding their intentions.

Although the number of studies on the neuropragmatics of language is still limited, there is a remarkable consistency in the finding that understanding the communicative intent of an utterance requires mentalizing. Since the linguistic code underdetermines speaker meaning, the ToM network needs to be invoked to get from coded meaning to speaker meaning. Despite the great popularity of the view that the Mirror Neuron System (MNS) is sufficient for action understanding (Rizzolatti and Sinigaglia, 2010), the MNS does not provide the crucial neural infrastructure for inferring speaker meaning. Next to core areas for retrieving lexical information from memory and unification of the lexical building blocks in producing and understanding multi-word utterances, other brain networks are needed to realize language-driven communication to its full extent (for a more extended review of studies on neuropragmatics, see Hagoort and Levinson, 2015).

Overall, we have seen that a number of neuronal networks are involved to get us from the spoken (or written) input to the unification of lexical elements into larger sentential and discourse configurations, and to extract the speaker meaning from the linguistic input. When speaking the recruitment of these networks mostly operates in the reversed temporal order. Some of the networks might show some domain-specificity for language, especially when it involves the coding of linguistic knowledge. Other networks, such as the one for cognitive control and the ToM network, are shared with other functional domains. A still largely open issue is how the interfaces and communication between these different networks is organized.

AcknowledgementS

I thank an anonymous reviewer for helpful comments. The content of this paper is inspired by my contributions in Hagoort (2005, 2013, 2014), Hagoort and Poeppel (2013), and Baggio and Hagoort (2011). The work was supported by the Spinoza Prize and the Academy Professorship Award, awarded to the author by the Netherlands Organization for Scientific Research (NWO) and the Netherlands Royal Academy for Arts and Sciences (KNAW), respectively.

References

Amodio, D.M., Frith, C.D., 2006. Meeting of minds: the medial frontal cortex and social cognition. Nat. Rev. Neurosci. 7, 268–277.
Anwander, A., Tittgemeyer, M., von Cramon, D.Y., Friederici, A.D., Knosche, T.R., 2007. Connectivity-based parcellation of Broca's area. Cereb. Cortex 17, 816–825.
Büring, D., 2007. Semantics, intonation and information structure. In: Ramchand, G., Reiss, C. (Eds.), The Oxford Handbook of Linguistic Interfaces. Oxford University Press, Oxford.
Bašnáková, J., Weber, K., Petersson, K.M., van Berkum, J., Hagoort, P., 2014. Beyond the language given: the neural correlates of inferring speaker meaning. Cereb. Cortex 24, 2572–2578.
Baggio, G., Hagoort, P., 2011. The balance between memory and unification in semantics: a dynamic account of the N400. Lang. Cognit. Proc. 26, 1338–1367.
Baggio, G., Van Lambalgen, M., Hagoort, P., 2008. Computing and recomputing discourse models: an ERP study. J. Mem. Lang. 59, 36–53.
Bastos, A.M., Usrey, W.M., Adams, R.A., Mangun, G.R., Fries, P., Friston, K.J., 2012. Canonical microcircuits for predictive coding. Neuron 76, 695–711.
Biswal, B.B., Kannurpatti, S.S., 2009. Resting-state functional connectivity in animal models: modulations by exsanguination. Methods Mol. Biol. 489, 255–274.
Biswal, B., Yetkin, F.Z., Haughton, V.M., Hyde, J.S., 1995. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. Magn. Reson. Imaging 34, 537–541.
Bock, K., 2011. How much correction of syntactic errors are there, anyway? Lang. Ling Compass 5, 322–335.
Bozic, M., Tyler, L.K., Ives, D.T., Randall, B., Marslen-Wilson, W.D., 2010. Bihemispheric foundations for human speech comprehension. PNAS 107, 17,439–17,444.
Brauer, J., Anwander, A., Friederici, A.D., 2011. Neuroanatomical prerequisites for language functions in the maturing brain. Cereb. Cortex 21, 459–466.
Brunel, N., Lavigne, F., 2009. Semantic priming in a cortical network model. J. Cognit. Neurosci. 21, 2300–2319.
Caplan, D., Waters, G.S., 1996. Syntactic processing in sentence comprehension under dual-task conditions in aphasic patients. Lang. Cognit. Proc. 11, 525–551.
Carrington, S.J., Bailey, A.J., 2009. Are there theory of mind regions in the brain? A review of the neuroimaging literature. Hum. Brain Mapp. 30, 2313–2335.
Catani, M., Allin, M.P.G., Husain, M., Pugliese, L., Mesulam, M.M., Murray, R.M., Jones, D.K., 2007. Symmetries in human brain language pathways correlate with verbal recall. PNAS 104, 17,163–17,168.
Chomsky, N., 1965. Aspects of the Theory of Syntax. The MIT Press, Cambridge, MA.
Cree, G.S., McRae, K., 2003. Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). J. Exp. Psychol. Gen. 132, 163–201.
Cree, G.S., McRae, K., McNorgan, C., 1999. An attractor model of lexical conceptual processing: simulating semantic priming. Cognit. Sci. 23, 371–414.
Davis, M.H., Coleman, M.R., Absalom, A.R., Rodd, J.M., Johnsrude, I.S., Matta, B.F., Menon, D.K., 2007. Dissociating speech perception and comprehension at reduced levels of awareness. PNAS 104, 16,032–16,037.
DeLong, K.A., Urbach, T.P., Kutas, M., 2005. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. Nat. Neurosci. 8, 1117–1121.
Deese, J., 1984. Thought into Speech: The Psychology of a Language. Prentice-Hall, Englewood Cliffs.
Dell, G.S., Chang, F., 2014. The P-chain: relating sentence production and its disorders to comprehension and acquisition. Phil. Trans. R. Soc. B 369, http://dx.doi.org/10.1098/rstb.2012.0394.
Durstewitz, D., Seamans, J.K., Sejnowski, T.J., 2000. Dopamine-mediated stabilization of delay-period activity in a network model of prefrontal cortex. J. Neurophysiol. 83, 1733–1750.
Erickson, T.D., Mattson, M.E., 1981. From words to meaning: a semantic illusion. J. Verbal Learn. Verbal Behav. 20, 540–551.
Ferreira, F., Bailey, G.D.K., Ferraro, V., 2002. Good-enough representations in language comprehension. Curr. Dir. Psycholog. Sci. 11, 11–15.
Fiorentino, R., Poeppel, D., 2007. Compound words and structure in the lexicon. Lang. Cognit. Proc. 22, 953–1000.
Fodor, J., Lepore, E., 2002. The Compositionality Papers. Oxford University Press, Oxford.
Frazier, L., 1987. Sentence processing: a tutorial review. In: Coltheart, M. (Ed.), Attention and Performance XII. Erlbaum, London, pp. 559–585.
Friederici, A.D., Ruschemeyer, S.A., Hahne, A., Fiebach, C.J., 2003. The role of left inferior frontal and superior temporal cortex in sentence comprehension: localizing syntactic and semantic processes. Cereb. Cortex 13, 170–177.
Friederici, A.D., 2009. Allocating functions to fiber tracts: facing its indirectness. Trends Cognit. Sci. 13, 370–371.
Friederici, A.D., 2012. The cortical language circuit: from auditory perception to sentence comprehension. Trends Cognit. Sci. 16, 262–268.
Friston, K.J., 2005. Hallucinations and perceptual inference. Behav. Brain Sci. 28, 764–766.
Fuster, J.M., 2009. Cortex and memory: emergence of a new paradigm. J. Cognit. Neurosci. 21, 2047–2072.
Gallistel, C.R., King, A.P., 2010. Memory and the Computational Brain: Why Cognitive Science Will Transform Neuroscience. Wiley-Blackwell, Chichester, West Sussex.
Garrido, M.I., Kilner, J.M., Kiebel, S.J., Friston, K.J., 2007. Evoked brain responses are generated by feedback loops. PNAS 104, 20,961–20,966.
Goucha, T., Friederici, A.D., 2015. The language skeleton after dissecting meaning: a functional segregation within Broca's Area. Neuroimage 114, 294–302.
Grainger, J., Rey, A., Dufau, S., 2008. Letter perception: from pixels to pandemonium. Trends Cognit. Sci. 12, 381–387.
Grice, P., 1989. Studies in the Way of Words. Harvard University Press, Cambridge, MA.
Hagoort, P., Indefrey, P., 2014. The neurobiology of language beyond single words. Annu. Rev. Neurosci. 37, 347–362.
Hagoort, P., Levinson, S.C., 2015. Neuropragmatics. In: Gazzaniga, M.S. (Ed.), The Cognitive Neurosciences. , 5 ed. MIT Press, Cambridge, Mass.
Hagoort, P., Poeppel, D., 2013. The infrastructure of the language-ready brain. In: Arbib, M.A. (Ed.), Language, Music, and the Brain: A Mysterious Relationship. MIT Press, Cambridge, MA, pp. 233–255.
Hagoort, P., Hald, L., Bastiaansen, M., Petersson, K.M., 2004. Integration of word meaning and world knowledge in language comprehension. Science 304, 438–441.
Hagoort, P., 2003. How the brain solves the binding problem for language: a neurocomputational model of syntactic processing. Neuroimage 20, S18–S29.
Hagoort, P., 2005. On Broca, brain: and binding: a new framework. Trends Cognit. Sci. 9, 416–423.
Hagoort, P., 2013. MUC (Memory, unification, control) and beyond. Front. Psychol. 4, Article 416.
Hagoort, P., 2014. Nodes and networks in the neural architecture for language: broca's region and beyond. Curr. Opin. Neurobiol. 28, 136–141.
Halliday, M.A.K., 1967. Notes on transitivity and theme in English. Part 2. J. Linguistics 3, 177–274.
Heim, I., Kratzer, A., 1998. Semantics in Generative Grammar. Blackwell, New York.
Hickok, G., Poeppel, D., 2004. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition 92S, 67–99.
Hickok, G., Poeppel, D., 2007. The cortical organization of speech processing. Nat. Rev. Neurosci. 8, 393–402.
Hoenig, K., Scheef, L., 2005. Mediotemporal contributions to semantic processing: fMRI evidence from ambiguity processing during semantic context verification. Hippocampus 15, 597–609.
Indefrey, P., Cutler, A., 2004. Prelexical and lexical processing in listening. In: Gazzaniga, M.S. (Ed.), The Cognitive Neurosciences III. MIT Press, Cambridge, MA, pp. 759–774.
Indefrey, P., Levelt, W.J., 2004. The spatial and temporal signatures of word production components. Cognition 92, 101–144.
Jackendoff, R., 1997. The Architecture of the Language Faculty. MIT Press, Cambridge, MA.
Jackendoff, R., 1999. The representational structures of the language faculty and their interactions. In: Brown, C.M., Hagoort, P. (Eds.), The Neurocognition of Language. Oxford University Press, Oxford, pp. 37–79.
Jackendoff, R., 2002. Foundations of Language. Oxford Univ. Press, New York.
Joshi, A.K., Schabes, Y., 1997. Tree-adjoining grammars. In: Rosenberg, G., Salomaa, A. (Eds.), Handbook of Formal Languages, Vol. 3. Springer-Verlag, New York, NY, pp. 69–123.
Keenan, E.L., 1979. On surface form and logical form. Stud. Ling. Sci. 8, 163–203.
Kelly, C., Uddin, L.Q., Shehzad, Z., Margulies, D.S., Castellanos, F.X., Milham, M.P., Petrides, M., 2010. Broca's region: linking human brain functional connectivity data and non-human primate tracing anatomy studies. Eur. J. Neurosci. 32, 383–398.
Kiebel, S.J., Daunizeau, J., Friston, K.J., 2008. A hierarchy of time-Scales and the brain. PLoS Comput. Biol. 4, e1000209.
Kiehl, K.A., Laurens, K.R., Liddle, P.F., 2002. Reading anomalous sentences: an event-related fMRI study of semantic processing. Neuroimage 17, 842–850.
Konopka, A.E., Meyer, A.S., 2014. Priming sentence planning. Cognit. Psychol. 73, 1–40.
Kotschi, T., 2006. Information structure in spoken discourse. In: Brown, K. (Ed.), Encyclopedia of Language & Linguistics. Elsevier, Oxford, pp. 677–683.
Kutas, M., Federmeier, K.D., 2011. Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). Annu. Rev. Psychol. 62, 621–647.
Kutas, M., Hillyard, S.A., 1980. Reading senseless sentences: brain potentials reflect semantic anomaly. Science 207, 203–205.
Lau, E., Stroud, C., Plesch, S., Phillips, C., 2006. The role of structural prediction in rapid syntactic analysis. Brain Lang. 98, 74–88.

Lau, E.F., Phillips, C., Poeppel, D., 2008. A cortical network for semantics: (de)constructing the N400. Nat. Rev. Neurosci. 9, 920–933.
Lebel, C., Beaulieu, C., 2009. Lateralization of the arcuate fasciculus from childhood to adulthood and its relation to cognitive abilities in children. Hum. Brain Mapp. 30, 3563–3573.
Levelt, W.J.M., 1989. Speaking: From Intention to Articulation. MIT Press, Cambridge, MA.
Levelt, W.J.M., 1999. Producing spoken language: a blueprint of the speaker. In: Brown, C.M., Hagoort, P. (Eds.), The Neurocognition of Language. Oxford University Press, Oxford, pp. 83–122.
Levinson, S.C., 2013. Action formation and ascription. In: Stivers, T., Sidnell, J. (Eds.), The Handbook of Conversation Analysis. Wiley-Blackwell, Malden, MA, pp. 103–130.
Liljeström, M., Hultén, A., Parkkonen, L., Salmelin, R., 2009. Comparing MEG and fMRI views to naming actions and objects. Hum. Brain Mapp. 30, 1845–1856.
Lisman, J.E., Fellous, J.M., Wang, X.J., 1998. A role for NMDA-receptor channels in working memory. Nat. Neurosci. 1, 273–276.
Müller, O., Hagoort, P., 2006. Access to lexical information in language comprehension: semantics before syntax. J. Cognit. Neurosci. 18, 84–96.
Mar, R.A., Oatley, K., 2008. The function of fiction is the abstraction and simulation of social experience. Perspect. Psychol. Sci. 3, 173–192.
Mar, R.A., 2011. The neural bases of social cognition and story comprehension. Annu. Rev. Psychol. 62, 103–134.
Mars, R.B., Sallet, J., Schuffelgen, U., Jbabdi, S., Toni, I., Rushworth, M.F., 2012. Connectivity-based subdivisions of the human right temporoparietal junction area: evidence for different areas participating in different cortical networks. Cereb. Cortex 22 (8), 1894–1903.
Marslen-Wilson, W.D., 1984. Function and process in spoken word-recognition. In: Bouma, H., Bouwhuis, D.G. (Eds.), Attention and Performance X: Control of Language Processes. Erlbaum, Hillsdale, NJ.
Marslen-Wilson, W.D., 2007. Morphological processes in language comprehension. In: Gareth, M. (Ed.), The Oxford Handbook of Psycholinguistics. Oxford University Press, Oxford, pp. 175–194.
Mason, R.A., Just, M.A., 2009. The role of the theory-of-Mind cortical network in the comprehension of narratives. Lang. Ling. Compass 3, 157–174.
Masson, M.E.J., 1991. A distributed memory model of context effects in word identification. In: Besner, D., Humphreys, G.W. (Eds.), Basic Processes in Reading: Visual Word Recognition. Erlbaum, Hillsdale, NJ, pp. 233–263.
Masson, M.E.J., 1995. A distributed memory model of semantic priming. J. Exp. Psychol. Learn. 21, 3–23.
McClelland, J., Elman, J., 1986. The TRACE model of speech perception. Cognitive Psychol. 18, 1–86.
McRae, K., Ross, B.H., 2004. Semantic Memory: Some Insights from Feature-based Connectionist Attractor Networks. Elsevier Academic Press, San Diego, CA.
Menenti, L., Gierhan, S.M.E., Segaert, K., Hagoort, P., 2011. Shared language: overlap and segregation of the neuronal infrastructure for speaking and listening revealed by functional MRI. Psychol. Sci. 22, 1173–1182.
Mesulam, M.-M., 1990. Large-scale neurocognitive networks and distributed processing for attention, language and memory. Ann. Neurol. 28, 597–613.
Mesulam, M.-M., 1998. From sensation to cognition. Brain 121, 1013–1052.
Miller, J., 2006. Focus. In: Brown, K. (Ed.), Encyclopedia of Language & Linguistics. Elsevier, Oxford.
Mitchell, J.P., Macrae, C.N., Banaji, M.R., 2006. Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. Neuron 50, 655–663.
Moss, H.E., Hare, M.L., Day, P., Tyler, L.K., 1994. A distributed memory model of the associative boost in semantic priming. Connect. Sci. 6, 413–427.
Nieuwland, M.S., Ditman, T., Kuperberg, G.R., 2010. On the incrementality of pragmatic processing: an ERP investigation of informativeness and pragmatic abilities. J. Mem. Lang. 63, 324–346.
Norris, D., 1994. Shortlist: a connectionist model of continuous speech recognition. Cognition 52, 189–234.
Partee, B.H., 1984. Compositionality. In: Veltman, F., Landmand, F. (Eds.), Varieties of Formal Semantics. Foris, Dordrecht.
Rilling, J.K., Glasser, M.F., Preuss, T.M., Ma, X., Zhao, T., Hu, X., Behrens, T.E., 2008. The evolution of the arcuate fasciculus revealed with comparative DTI. Nat. Neurosci. 11, 426–428.
Rizzolatti, G., Sinigaglia, C., 2010. The functional role of the parieto-frontal mirror circuit: interpretations and misinterpretations. Nat. Rev. Neurosci. 11, 264–274.
Rodd, J.M., Davis, M.H., Johnsrude, I.S., 2005. The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity. Cereb. Cortex 15, 1261–1269.
Rodd, J.M., Longe, O.A., Randall, B., Tyler, L.K., 2010. The functional organisation of the fronto-temporal language system: evidence from syntactic and semantic ambiguity. Neuropsychol 48, 1324–1335.
Ruschemeyer, S.A., Zysset, S., Friederici, A.D., 2006. Native and non-native reading of sentences: an fMRI experiment. Neuroimage 31, 354–365.
Saxe, R., Moran, J.M., Scholz, J., Gabrieli, J., 2006. Overlapping and non-overlapping brain regions for theory of mind and self reflection in individual subjects. Soc. Cogn. Affect. Neurosci. 1, 229–234.
Saxe, R., 2010. The right-temporo-parietal junction: a specific brain region for thinking about thoughts. In: Leslie, A., German, T. (Eds.), Handbook of Theory of Mind. Psychology Press, Philadelphia, PA.
Seuren, P.A.M., 1985. Discourse Semantics. Basil Blackwell, Oxford.
Sherman, S.M., Guillery, R.W., 1998. On the actions that one nerve cell can have on another: distinguishing drivers from modulators. PNAS 95, 7121–7126.
Sherman, S.M., 2007. The thalamus is more than just a relay. Curr. Opin. Neurobiol. 17, 417–422.
Snijders, T.M., Vosse, T., Kempen, G., Van Berkum, J.J.A., Petersson, K.M., Hagoort, P., 2009. Retrieval and unification of syntactic structure in sentence comprehension: an fMRI study using word-category ambiguity. Cereb. Cortex 19, 1493–1503.
Solomyak, O., Marantz, A., 2010. Evidence for early morphological decomposition in visual word recognition. J. Cognit. Neurosci. 22, 2042–2057.
Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M., Sedivy, J.C., 1995. Integration of visual and linguistic information in spoken language comprehension. Science 268, 1632–1634.
Thompson-Schill, S.L., Bedny, M., Goldberg, R.F., 2005. The frontal lobes and the regulation of mental activity. Curr. Opin. Neurobiol. 15, 219–224.
Thorpe, S., Fize, D., Marlot, C., 1996. Speed of processing in the human visual system. Nature 381, 520–522.
Tyler, L.K., Marslen-Wilson, W.D., Randall, B., Wright, P., Devereux, B.J., Zhuang, J., Stamatakis, E.A., 2011. Left inferior frontal cortex and syntax: function: structure and behaviour in left-hemisphere damaged patients. Brain 134, 415–431.
Van Berkum, J.J.A., Brown, C.M., Zwitserlood, P., Kooijman, V., Hagoort, P., 2005. Anticipating upcoming words in discourse: evidence from ERPs and reading times. J. Exp. Psychol. Learn. 31, 443–467.
Van Berkum, J.J.A., Van den Brink, D., Tesink, C., Kos, M., Hagoort, P., 2008. The neural integration of speaker and message. J. Cognit. Neurosci. 20, 580–591.
Vigneau, M., Beaucousin, V., Herve, P.Y., Jobard, G., Petit, L., Crivello, F., Tzourio-Mazoyer, N., 2010. What is right-hemisphere contribution to phonological, lexico-semantic, and sentence processing? Insights from a meta-analysis. Neuroimage 54, 577–593.
Vosse, T., Kempen, G.A.M., 2000. Syntactic structure assembly in human parsing: a computational model based on competitive inhibition and lexicalist grammar. Cognition 75, 105–143.
Wang, L., Bastiaansen, M., Yang, Y., Hagoort, P., 2011. The influence of information structure on the depth of semantic processing: how focus and pitch accent determine the size of the N400 effect. Neuropsychology 49, 813–820.
Wang, X.-J., 1999. Synaptic basis of cortical persistent activity: the importance of NMDA receptors to working memory. J. Neurosci. 19, 9587–9603.
Willems, R.M., Özyürek, A., Hagoort, P., 2007. When language meets action: the neural integration of gesture and speech. Cereb. Cortex 17, 2322–2333.
Willems, R.M., Özyürek, A., Hagoort, P., 2008. Seeing and hearing meaning: event-related potential and functional magnetic resonance imaging evidence of word versus picture integration into a sentence context. J. Cognit. Neurosci. 20, 1235–1249.
Wong, K.-F., Wang, X.-J., 2006. A recurrent network mechanism of time integration in perceptual decisions. J. Neurosci. 26, 1314–1328.
Xiang, H., Fonteijn, H.M., Norris, D.G., Hagoort, P., 2010. Topographical functional connectivity pattern in the Perisylvian language networks. Cereb. Cortex 20, 549–560.
Zwitserlood, P., 1989. The locus of the effects of sentential-semantic context in spoken-word processing. Cognition 32, 25–64.