



## From Multimodal Sensorimotor Integration to Semantic Networks: A Phylogenetic Perspective on Speech and Language Evolution

Maëva Michon<sup>1,2</sup> and Francisco Aboitiz<sup>2</sup>

<sup>1</sup>Praxiling Laboratory, UMR 5267, CNRS, Université Paul Valéry, Montpellier, France

<sup>2</sup>Laboratory for Cognitive and Evolutionary Neuroscience, Interdisciplinary Center for Neuroscience, Department of Psychiatry, Faculty of Medicine, Pontificia Universidad Católica de Chile, Santiago, Chile

**Keywords:** language evolution, multimodal integration, perisylvian regions, semantic network, sensorimotor, superior temporal sulcus (STS)

### ABSTRACT

This integrative perspective article delves into the crucial role of the superior temporal sulcus (STS) and adjacent perisylvian regions in multimodal integration and semantic cognition. Drawing from a wide range of neuroscientific evidence, including studies on nonhuman primates and human brain evolution, the article highlights the significance of the STS in linking auditory and visual modalities, particularly in the establishment of associative links between auditory inputs and visual stimuli. Furthermore, it explores the expansion of the human temporal lobe and its implications for the amplification of multisensory regions, emphasizing the role of these regions in the development of word-related concepts and semantic networks. We propose a posteroanterior gradient organization in the human temporal lobe, from low-level sensorimotor integration in posterior regions to higher-order, transmodal semantic control in anterior portions, particularly in the anterior temporal lobe. Overall, this perspective provides a comprehensive overview of the functional and evolutionary aspects of the STS and adjacent regions in multimodal integration and semantic cognition, offering valuable insights for future research in this field.

**Citation:** Michon, M., & Aboitiz, F. (2025). From multimodal sensorimotor integration to semantic networks: A phylogenetic perspective on speech and language evolution. *Neurobiology of Language*, 6, nol\_a\_00164. [https://doi.org/10.1162/nol\\_a\\_00164](https://doi.org/10.1162/nol_a_00164)

**DOI:**  
[https://doi.org/10.1162/nol\\_a\\_00164](https://doi.org/10.1162/nol_a_00164)

**Received:** 10 January 2024  
**Accepted:** 3 February 2025

**Competing Interests:** The authors have declared that no competing interests exist.

**Corresponding Authors:**  
Francisco Aboitiz  
[fabotiz@uc.cl](mailto:fabotiz@uc.cl)  
Maëva Michon  
[maeva.michon@hotmail.fr](mailto:maeva.michon@hotmail.fr)

**Handling Editor:**  
Steven Small

---

**Copyright:** © 2025  
Massachusetts Institute of Technology  
Published under a Creative Commons  
Attribution 4.0 International  
(CC BY 4.0) license



The MIT Press

### INTRODUCTION

Traditionally, the research on sensory neuroscience has long been divided along modality-specific lines, with experts in the field of auditory or visual or motor systems. This dominant approach has led to a focus on questions concerning specific primary sensory pathways. Not until more recently did a field of research on multisensory neuroscience emerge and consolidate (Stein et al., 2020). In the past, our understanding of how we perceive objects was primarily grounded in visual processing alone. The neuroscience of vision revealed that to achieve object identification and localization, different types of information are processed in different neural pathways. Audition neuroscientists have then shown that objects can also be perceived auditorily with dedicated brain pathways to localize and identify sounds. It seems evident that, additionally to shape and sound (i.e., visual and auditory information) different features, like weight, texture, and even taste or smell, can be used to perceive individual objects, not just their shape. These multisensory representations probably enhance object perception, making it more reliable (Newell et al., 2023). Importantly, at the crossroads between seemingly sensory-specific pathways, subcortical and cortical areas with multisensory

response profiles have been described. For instance, neurons in the superior colliculus and temporal sulcus display responses to multisensory stimuli that exceed the linear sum of individual responses to sensory-specific stimuli presented separately. The discovery of such a property, known as super-additivity, has had critical implications for the neuroscience of multisensory integration (Stein & Meredith, 1993).

The capacity of neurons to integrate multisensory information is not innate. Instead, it develops with repeated exposure to pieces of sensory information perceived concomitantly and accumulating substantial cross-modal experience (Bean et al., 2023; Smyre et al., 2024). Our interactions with the world are multisensorial and our conceptual knowledge about it requires cross-modal (visual, auditory, tactile, gustatory, somatosensory, and motor) integration (Lambon Ralph, 2014), resulting in a meaningful, adaptative, and coherent perception of the world (Gomez-Marin & Ghazanfar, 2019; Varela et al., 2017). Recent research has suggested that the representation of concepts is largely rooted in our experiential knowledge, involving both action and perception (Fernandino et al., 2016; Kuhnke, Kiefer, & Hartwigsen, 2023; Pouw et al., 2021; Tong et al., 2022). Congruently, there is increasing support for the idea that sensorimotor representations upon which concepts were acquired are reactivated for the comprehension of words conveying concepts (Fernandino & Conant, 2023; Fernandino et al., 2022; Kuhnke et al., 2020). Based on statistical associative learning (i.e., the detection of statistical regularities such as co-occurrences between stimuli in our environment), the human brain is able to generate multisensory predictions. For example, young children can associate the word “dog” with the shape of a dog and the barking of a dog.

Multimodal:  
Requiring the involvement of  
multiple sensory modalities.

This perspective article aims to provide a comprehensive overview of the functional and evolutionary aspects of the superior temporal sulcus (STS) and perisylvian regions in multimodal integration and semantic cognition. In the next section, we highlight the role of sensory and motor systems in conceptual knowledge, suggesting that modality-specific sensorimotor brain regions and multimodal convergence zones are engaged to give meaning to the world. A Multimodal Interface Along STS addresses the similarities in organizational structures and processing pathways between the auditory and visual systems, emphasizing the integration of multimodal sensory information in the STS and temporoparietal regions. This multimodal interface is crucial for understanding how the brain processes and integrates sensory information within the context of social interaction and communication. In Specialization of the pSTS During Brain Evolution, we examine the possible mechanisms by which this multimodal interface in the temporal lobe has consolidated during brain evolution, allowing the emergence of a proto-language in our ancestors. We and others have proposed that the expansion of the cortical regions associated with the STS facilitated the generation of multimodal representations including sound-object associations, amplifying multimodal representations that provided an early scaffolding for the emergence of semantic links to proto-linguistic utterances, that is, names for objects (Aboitiz, 2018a, 2018b; Kausel et al., 2024; Petrides, 2023). Finally, we discuss the anatomo-functional and phylogenetic constraints of the *Homo* lineage which may have enabled our ancestors to develop an increasing vocal repertoire, leading to the development of early human language.

#### **From Representations to Reenactment: The Re-Engagement of Multimodal Experiences**

During oral communication, understanding the meaning of a word is thought to imply an association between perceived speech sounds forming words and the physical aspects of the referred object, person, or action afforded by our sensory systems. Objects, people, or action

recognition can be performed via different sensory modalities. A person can be recognized by her face or voice because we have learned to associate the identity of this person, the sound of her name with the sound of her voice and how she looks (audiovisual associations). Fruits can be recognized by associating their names with their shapes, colors, tastes and smells (Lambon Ralph, 2014). Across languages, words (particularly nouns and action verbs) are related to different modalities of sensory experience, and their representations are multimodal (Ibáñez et al., 2023; Lynott et al., 2020; Miklashevsky, 2018; Speed & Brysbaert, 2024; Speed & Majid, 2017; Vergallito et al., 2020; Zhong et al., 2022). A recent neuroimaging study has evidenced a distributed network for multimodal experiential representation of concepts and showed that “the retrieval of conceptual knowledge during word comprehension relies on a much larger portion of the cerebral cortex than previously thought and that multimodal experiential information is represented throughout the entire network” (Tong et al., 2022, p. 7121). This implies that linguistic concepts rely on multimodal, possibly associative networks distributed across different cortical regions (Lambon Ralph et al., 2017).

**Reenactment:**  
Bringing past experiential knowledge  
to the present.

**Active inference:**  
Learning by minimizing the  
difference between predictions and  
sensory inputs through actions and  
belief updates.

Rather than understanding meaning as the processing of amodal, encapsulated representations, an increasing number of scientists is advocating for the grounding of meaning in the reenactment of these multimodal experiences with the world (Borghi et al., 2024; Calzavarini, 2024; Dove, 2023a; Kewenig et al., 2024; Pulvermüller, 2013). This approach, in our opinion, is biologically more plausible because living organisms learn about and understand the world by purposefully interacting with their environment and predicting the outcomes of those interactions. This process helps them to develop a fundamental understanding and a sense of significance, which becomes the ground for their subsequent knowledge (Pezzulo et al., 2024). From this standpoint, the living organism is not a passive observer of an external, predetermined reality but rather a proactive agent that constructs its own reality by interacting with its environment and predicting the sensory consequences of its action. A notion compatible with this view, called *active inference*, proposes that sensory and motor systems are actively recruited during perception (Parr et al., 2022; Ramstead et al., 2020; Schroeder et al., 2010). Through repeated interactions with its environment, an organism can identify regularities and extract sensorimotor contingencies. Importantly, the organism interplays with the environment through its actions and eventually becomes able to predict the changes its own or others’ actions can produce in the somatosensory input. According to this framework, the semantic content of a concept is achieved by reusing the perceptual and motor areas of the brain that allow us to sense and interact with the world (Dove, 2023b, 2024; Pulvermüller, 2013; Pulvermüller & Fadiga, 2010).

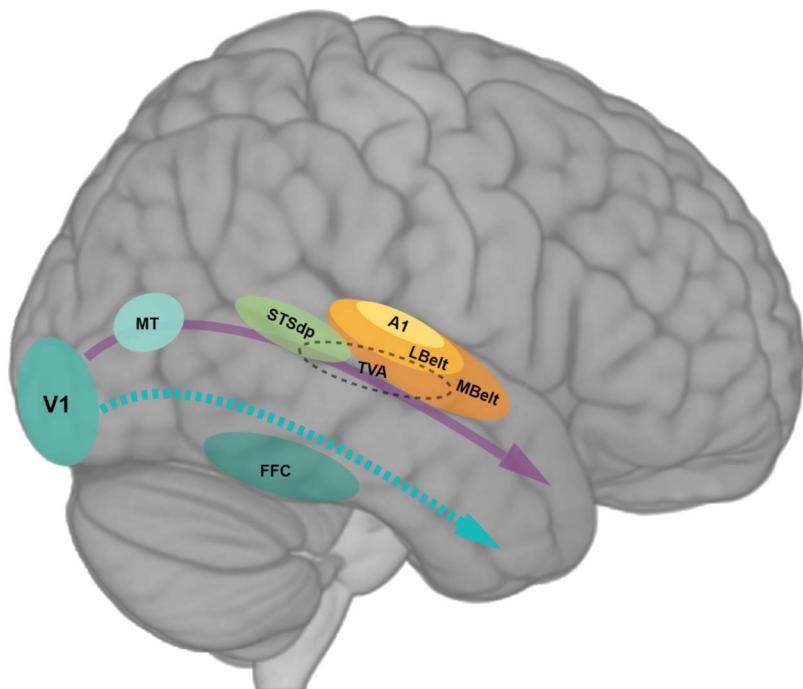
Importantly, the grounding of conceptual knowledge does not seem to be limited to concrete concepts. Using functional magnetic resonance imaging (fMRI), Harpaintner and collaborators (2020) found that processing abstract concepts with a known motor (e.g., effort) versus visual (e.g., beauty) feature content respectively activates action-related and vision-related brain regions, as identified in single subjects’ brains by functional localizer tasks (hand movement for the motor localizer and picture exploration for the visual localizer). A recent meta-analysis of 212 neuroimaging studies reported that “conceptual processing consistently engages brain regions also activated during real perceptual-motor experience of the same modalities” (Kuhnke, Beaupain, et al., 2023, abstract). The authors propose a novel hierarchical model for conceptual knowledge engaging not only modality-specific sensorimotor brain regions but also multimodal convergence zones, such as the inferior parietal lobe and the posterior middle temporal gyrus. In the remaining sections of this perspective article, we address the nature and functional relevance of these multimodal convergence zones for speech and language with greater detail.

### A Multimodal Interface Along STS

In their inaugural model, Mortimer Mishkin and Leslie Ungerleider (1982) initially outlined a dual pathway arrangement of the visual system in the nonhuman primate brain. This organization featured a ventral stream specialized in object recognition and a dorsal stream dedicated to determining objects' spatial location. Ten years later, a comparable arrangement was described in the human visual system (Goodale & Milner, 1992). This anatomo-functional organization was then expanded to the auditory system of both nonhuman (Kaas & Hackett, 1999; Rauschecker & Tian, 2000) and human primates (Arnott et al., 2004), highlighting a ventral "what" and a dorsal "where" stream specialized for sound recognition and spatial localization, respectively. The description of this functional organization in "low-level" sensory cortices has subsequently inspired dual-pathway theories for "higher-order" cognitive functions, such as attention, exemplified by the ventral and dorsal attentional network proposed by Corbetta and Shulman (2002), and language, illustrated by the ventral and dorsal streams for language posited by Hickok and Poeppel (2007). Nearly 40 years after Mishkin and Ungerleider's (1982) original proposition, however, there is growing evidence indicating the need for an update to the dual-stream model of the visual pathway.

### The Visual System

It was recently proposed that the dual stream model for the organization of the visual system should be updated to include a third, lateral pathway specialized for dynamic social perception (Pitcher, 2021; Pitcher & Ungerleider, 2021). The authors argue for a third stream anatomically and functionally independent from the ventral stream (see purple and blue arrows in Figure 1), which projects on the lateral surface of both humans' and macaques' brains, from primary visual cortex (V1) to anterior STS (aSTS), encompassing visual motion visual middle temporal area (V5/MT), extrastriate body area (EBA) and the posterior STS (pSTS). This third visual stream is thought to be involved during the processing of socially relevant dynamic biological motion such as body posture and facial expression and to support, along with other regions of the temporoparietal union, higher-order functions like social cognition and theory of mind. Recently, using transcranial magnetic stimulation, Gandolfo et al. (2024) reported the causal involvement of the left EBA in detecting visual signals of social interaction. Wurm and Caramazza (2022), on the other hand, have argued for a subdivision of the ventral "what" pathway into two streams; a ventral and a lateral stream specialized for object and action recognition, respectively. Interestingly, and in line with the proposal of a third visual stream "specialized for interpreting the physical actions that we use to understand others" (Pitcher, 2023, p. R1222), Wurm and Caramazza (2022) described a dorsoventral gradient in the organization of the lateral occipitotemporal cortex, with dorsal portions responding to animated entities and ventral portions to unanimated entities. They also propose a second dimension of organization of this lateral action stream along a posterior–anterior gradient from visual perceptual precursors to more specific semantic representations of dynamic actions and objects. Similar organizational principles of the visual system were formerly described by Weiner and Grill-Spector (2013), who proposed "a new model of high-level visual cortex consisting of ventral, lateral, and dorsal components, where multimodal processing related to vision, action, haptics, and language converges in the lateral pathway" (p. 74) with potential implications for social communication. More recently, McMahon and collaborators further explored the involvement and relevance of this lateral visual pathway for social interactions. Congruently with other proposals addressed above, they claim that the lateral visual pathway for social action recognition is hierarchically organized from posterior to



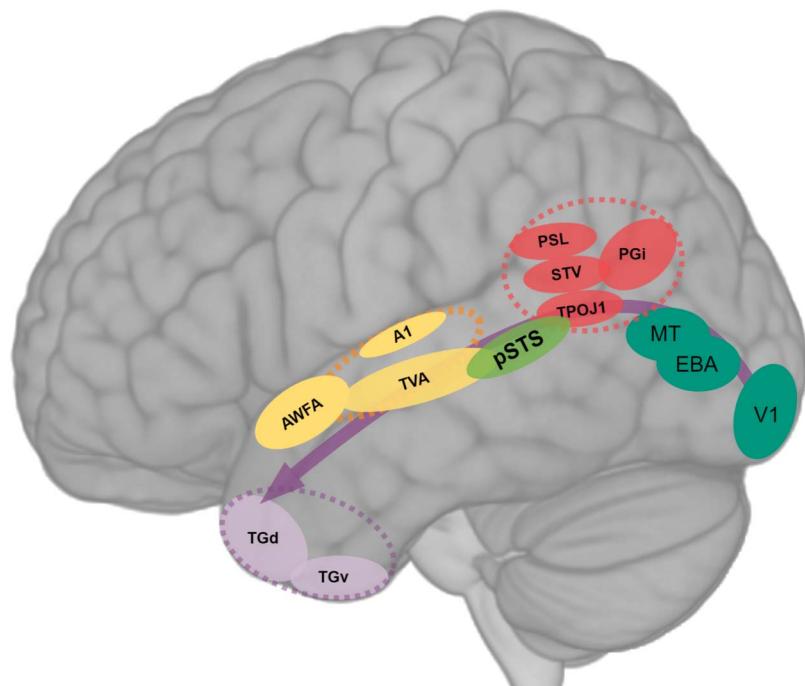
**Figure 1.** Right STSdp as an interface for multimodal integration of social stimuli. Sensory-specific systems process visual and auditory social stimuli such as faces and voices in the occipital and temporal gyri. Along the lateral surface of the brain, at the interface of auditory (yellow and orange areas) and visual (blue areas) regions, stands STSdp (green area), which is increasingly accepted as an area supporting multisensory integration. The lateral visual stream (solid purple arrow) is anatomically and functionally different from the ventral visual stream (dashed blue arrow), the former dealing with dynamic, animated aspects of social perception and action recognition while the latter seems to respond to static or unanimated objects and people. (Areas depicted by solid colored circles are named after the Human Connectome Project nomenclature). V1 = primary visual cortex; FFC = fusiform face cortex; MT = middle temporal area; A1 = primary auditory cortex; LBelt = lateral auditory belt; MBelt = medial auditory belt; TVA = temporal voice area; STSdp = superior temporal sulcus dorsal posterior.

anterior regions, with high-level social interaction information being processed along the STS (McMahon et al., 2023; McMahon & Isik, 2023, 2024).

### The Auditory System

Spoken words are produced by a specific sequence of movements executed by the speaker's orofacial articulators (aka *articulemes*, Michon et al., 2019, 2020), which, once uttered, is perceived as a voice conveying a short stream of speech sounds in the listener's brain. Extending bilaterally from the pSTS to the middle and anterior superior temporal gyri (mSTG and aSTG) stands an area known as the temporal voice area (TVA; depicted as a dashed ellipse in Figure 1). Compared to the auditory belts (depicted as yellow and orange ellipses in Figure 1), which are dedicated to the processing of general sound features, the TVA shows specific responsiveness to conspecific vocalizations over nonvocal sounds in the human brain (Agus et al., 2017; Belin et al., 2000; Bodin et al., 2021; Pernet et al., 2015). Single electrode recordings of neural activity in the human STG have revealed response selectivity to phonetic and acoustic features of continuous speech (Mesgarani et al., 2014). In non-human primates' brains, neurons that respond selectively to species-specific vocalizations were previously identified in the anterior (or rostral) auditory belts of rhesus macaques (Rauschecker et al., 1995).

Recently, similar regions for vocalization processing have been described in marmosets' temporo-frontal cortices, suggesting that the voice perception circuit in the human brain may have evolved from a precursor circuit for vocalization processing predating the separation of the Old and New World primates (Belin et al., 2023; Jafari et al., 2023). It is noteworthy that the TVA at least partially overlaps with anterior auditory belt areas (depicted in dashed yellow circle in Figure 2), which were meticulously mapped with single-unit techniques in rhesus monkeys by Tian and collaborators (2001), and defined as functionally specialized for species-specific communication calls. More anteriorly, in the human ventral auditory stream, an auditory word form area was discovered in the left aSTG and called after its function for eliciting highly selective activity in response to spoken words, as opposed to written words which are processed in an area of the ventral visual stream called the visual word form area in the left fusiform gyrus (DeWitt & Rauschecker, 2012; Rolls et al., 2023). Reminiscent of the organizational principle of the visual ventral and lateral pathways, the ventral auditory stream seems to process increasingly complex vocal sounds (from phonemes to words to phrases) when moving forward to the anterior regions of the superior temporal lobe. Accordingly, new evidence suggests convergent functional organization strategies are present across modalities in the auditory and visual ventral streams (Damera et al., 2023).



**Figure 2.** Evolution of a multimodal network for semantics. This neural circuit evolved from ancestral networks sustaining sensorimotor associations between modality-specific information to an expanded network enabling the mapping of linguistic labels (phonological and articulatory features of words and names) with multimodal information concerning the referred object, action or person. This network extends along the superior temporal sulcus and its progressive expansion in the left hemisphere allowed the consolidation of a proto-lexicon and the corresponding semantic relations with the empiric world. Dashed red circle indicates temporoparietal junction; dashed yellow circle = lateral auditory belt and medial auditory belt; dashed purple circle = anterior temporal lobe. V1 = primary visual cortex; MT = middle temporal area; EBA = extrastriate body area; PSL = perisylvian language area; STV = superior temporal visual area; PGi = inferior parietal PG area; TPOJ1 = temporoparietooccipital junction 1; pSTS = posterior superior temporal sulcus; A1 = primary auditory cortex; TVA = temporal voice area; AWFA = auditory word form area; TGd = lateral temporal TG dorsal; TGv = lateral temporal TG ventral.

### **Multimodal Integration at the Crossroads of Visual and Auditory Systems**

While distinct in their primary functions, the auditory and visual systems exhibit remarkable similarities in their organizational structures and processing pathways. Both systems are organized along a posteroanterior gradient, beginning with low-level perceptual features and advancing to anterior regions responsible for more complex semantic representations, such as speech sound processing in the auditory stream (Leaver & Rauschecker, 2010; Norman-Haignere et al., 2015) and object and action recognition in the visual stream (Wurm & Caramazza, 2022). This parallel organization highlights a convergent evolutionary strategy in the brain's handling of different sensory modalities. Furthermore, recent research suggests that these sensory systems are not entirely isolated; they interact and integrate information, particularly in regions like the superior temporal lobe, which serves as a crucial interface. This interconnectedness is pivotal for understanding how the brain processes and integrates multimodal sensory information, setting the stage for examining the detailed pathways and their specialized functions within the context of social interaction and communication.

During social interactions, the faces and voices of interlocutors are often perceived together, simultaneously providing congruent visual and auditory communicative cues. In addition to the fusiform face and the occipital face areas in the ventral visual stream, STS dorsal posterior (STSdp) is also considered a classical core region for face processing. According to the evidence reported above in *The Visual System*, STSdp is involved particularly in response to dynamic as compared to static face processing, such as facial expressions but also visual speech perception. It is noteworthy that the STSdp receives inputs from both the ventral auditory stream and the lateral visual stream and is considered a hub for multimodal integration (Landsiedel & Koldewyn, 2023; Rolls et al., 2023). In nonhuman primates studies using single-unit recordings of brain activity, audiovisual integration has been reported in macaque face patch anterior fundus in STS, but not anterior medial in the anterior ventral temporal cortex (Ghazanfar et al., 2005; Khandhadia et al., 2021; Perrodin et al., 2014). Using fMRI in humans, Zhu and Beauchamp (2017) reported that the face-sensitive pSTS contains different neuronal populations that respond preferentially to eye movements or mouth movements and showed that only the subregion preferring mouth movements is also strongly responsive to voices (also see Rennig & Beauchamp, 2018). Congruent with the functional properties of this pSTS subregion, an area that is selective for visual speech versus nonspeech mouth movements has been identified in the left hemisphere. According to its function and location at the boundary of pMTG and pSTS, this area has been labeled the temporal visual speech area (TVSA; Bernstein et al., 2011; Bernstein & Liebenthal, 2014). Recently, using inhibitory transcranial magnetic stimulation, Jeschke et al. (2023) demonstrated the causal influences of area V5/MT, adjacent to the TVSA, on visual speech recognition. Importantly, as Rauschecker and Afsahi (2023) outlined, "STS regions also receive visual inputs about moving faces and objects, and the auditory and visual streams are combined to help in multimodal object identification, such as who is speaking, what is being said, what the object is, and so on" (Rauschecker & Afsahi, 2023, p. 1888). This connectivity pattern suggests that beyond its involvement in multimodal speech perception (Kausel et al., 2024), STS may also be crucial for grounded semantic cognition.

### **Specialization of the pSTS During Brain Evolution**

#### ***From action recognition to social cognition***

Dorsal to area pSTS, the temporoparietal junction (TPJ) is largely accepted to be a key region of the social cognition network in both humans and nonhuman primates. Often considered as an anatomically ill-defined region, TPJ is a term used to refer to a set of cortical areas including

the pSTS and parts of the inferior parietal lobe, namely the angular and supramarginal gyri. It receives and integrates information from the visual, auditory, somatosensory systems as well as the thalamus and the limbic system. It is known, for instance, for its role in the processing of emotional facial expressions and theory of mind operations (Roumazeilles et al., 2021). Interestingly, Patel et al. (2019) compared humans' and macaques' brains and found that the pSTS-TPJ region has expanded and rearranged along brain evolution. These authors proposed that anatomical and functional reorganization of pSTS-TPJ has served as a basis for the consolidation of a third processing stream that allowed the emergence of increased social abilities in humans. pSTS-TPJ is conceived as a hub that processes facial and other biological motion information allowing us to generate internal models of social scenarios based on previous empirical and multisensory social interactions. Somehow, this area bridges together the information required for understanding the mental states (e.g., intentions, beliefs, desires) of others and thus for the emergence of meaning about our social environment. It is also involved in our ability to imitate, which mainly relies on cross-modal associations of sensorimotor information (e.g., visuomotor associations for facial or gestural imitation and audiomotor associations for vocal imitation) and is crucial to social adaptation and learning (Michon et al., 2022).

FMRI studies in humans have reported that brain activity in response to the integration of people-related (vs. objects-related) information, particularly faces and voices, is restricted to the right pSTS, suggesting a functional lateralization of this area (Landsiedel & Koldewyn, 2023; Watson et al., 2014). A structural asymmetry of the depth of STS was also reported in human newborns, infants, children and adults, this sulcus being deeper in the right than in the left hemisphere (Bodin et al., 2018; Dubois et al., 2010; Leroy et al., 2015). Since they only observed a barely visible asymmetry in macaques, Leroy et al. (2015) argued in favor of the genetic origin of this human-specific STS asymmetry that could have favored the evolution of increasingly sophisticated communicative and social cognition abilities. However, Hopkins et al. (2023) recently reported a small but significant rightward asymmetry in STS depth in a sample of 292 chimpanzees, challenging the idea that this asymmetry pattern is uniquely human. Like humans, chimpanzees also use multimodal signals to communicate. To understand each other, they need to integrate their conspecifics' facial, vocal, and gestural behaviors. The right STS is known for its fundamental role in interpreting the meaning and communicative intentions of these multimodal stimuli (Redcay, 2008). The presence of this rightward asymmetry in STS depth in chimpanzees (although smaller compared to humans) is therefore more suggestive of gradual and continuous structural changes along nonhuman primate brain evolution. It is noteworthy that nonhuman primates living in larger social groups show an increased volume of white matter (Meguerditchian et al., 2021) and gray matter, especially in the mid-STS and prefrontal regions (Sallet et al., 2011). A deeper understanding of the evolutionary trajectory in the degree of pSTS asymmetry and the relation between its depth and prosocial behaviors in nonhuman primates have the promising potential to shed new light on the origin of our social brains.

#### ***From the grounding of meaning to semantic networks***

Evidence supporting the existence of a third visual pathway for social perception comes from studies of both human and nonhuman primate brains. Interestingly, Pitcher and Ungerleider (2021) raised the possibility of lateralization of the third visual pathway in the right hemisphere, leaving the question of the potential role of this stream in the left hemisphere unanswered. Different populations of neurons within the left pSTS were identified that fired in response to mouth or eye movements. Remarkably, those neurons that fired for mouth (but not for eyes) movements also responded to conspecific voices (Rennig & Beauchamp,

2018, 2022; Zhu & Beauchamp, 2017). Similar multisensory processing of visual and auditory communicative signals has been reported in rhesus macaques (Froesel et al., 2022; Ghazanfar et al., 2005, 2008). As proposed elsewhere (Aboitiz, 2018a, 2018b), due to its privileged projection along the visual system processing dynamic faces and the auditory system processing voices, the third visual stream for social perception in the language-dominant hemisphere represents a well-suited neural network to support audiovisual integration of human speech (Kausel et al., 2024; Michon et al., 2024) and nonhuman primates' lip-smacking for communicative purposes (Michon et al., 2022).

Here, we propose that in addition to its involvement in multimodal speech perception in our contemporary human brains, the left pSTS has played a critical role in the emergence of meaningful communicative behaviors and the consolidation of a proto-lexicon and primitive semantics in our ancestors' brains. As a multimodal region at the interface of the ventral auditory and visual pathways, pSTS has strengthened the referring-referred mapping, namely associations between acoustic properties of conspecific vocalizations produced in a particular context (e.g., alarm call) and the visual information about surrounding objects, actions, or events (e.g., the presence of a predator). The medial temporal gyrus (MTG) has progressively appeared and expanded along brain evolution (Roumazeilles et al., 2020) and its strengthening projections toward frontal articulatory regions via the arcuate fasciculus in humans' left hemisphere have facilitated the transformation of phonological information into speech motor commands (Aboitiz, 2018a; Skeide & Friederici, 2016), allowing the reproduction or imitation of intentional communicative vocalizations within a given social group. Then, learned audio-visual-motor associations between vocalizations and visual representations along the pSTS may have contributed to the apparition of a primitive lexicon providing meaning to phonological and articulatory sequences. With increasingly organized societies, this proto-lexicon became more complex involving multisensory information from our environment and eventually abstract concepts. The expansion of this lexicon may have increased the exposure to conventionalized articulatory sequences consolidated by phonological memory and allowed the use of vocalizations to refer to objects, people, or events even in their absence in the visual scene.

In line with our proposal, Petrides (2023) recently argued in favor of a polysensory neural population in the left STS that integrates highly processed visual information and highly processed auditory information. He claimed that semantic processing requires both sensory-specific information and multisensory integration of information and that our ability for semantics has evolved from the ventral expansion of the polysensory STS to form the adjacent MTG.

## DISCUSSION

In earlier works, we have proposed the hypothesis that in the *Homo* lineage, the growth of auditory-vocal connectivity via an expansion of the arcuate fasciculus and related tracts of the dorsal auditory pathway enabled our ancestors with an increasing vocal repertoire (Aboitiz, 2018a; Aboitiz & García, 1997; see also Rilling et al., 2008; Scott et al., 2012). In light of this evidence, we suggest that the increasing plasticity of vocal (and manual) behavior in hominins was accompanied by the development of an amplified network in which different sensorimotor components (particularly visuomotor and auditory-vocal) interacted to provide the early meanings that led to early human language, possibly using a combination of gestures and learned vocal calls. In this article, we develop an argument highlighting the co-option of the multisensory STS, the anterior temporal lobe (ATL), and the TPJ to support an interface "translating" linguistic elements like words or proto-words into a multimodal network depicting real objects or events (see also Aboitiz, 2018b).

As mentioned, several previous reports have highlighted the role of the STS and the ATL in multimodal integration, working as an interface between the ventral auditory stream in the STG and the ventral visual stream along the inferior temporal gyrus (see Binder & Desai, 2011; Binder et al., 2009; Perrodin et al., 2015; Petrides, 2014). Compared to nonhuman primate, the human temporal lobe has expanded, yielding the adjacent MTG (Roumazeilles et al., 2020; Sierpowska et al., 2022), which represents an amplification of this multisensory region. In this line, some authors have proposed that in the left hemisphere, this region became critical for establishing associative links between auditory inputs and visual stimuli, which may have been critical for social communication in early humans (Petrides, 2014, 2023). Together with gestural communication (Becker & Meguerditchian, 2022; Meguerditchian, 2022), the STS and adjacent regions may have provided a primitive scaffolding linking vocalizations or gestures with visual and other modalities. (Recall that neurons in the macaque STS may respond to auditory, visual, somesthetic, or a combination of these stimuli; Bruce et al., 1981.) Consistent with this notion, in non-human primates, complex visual and social input to the STS works by enhancing or inhibiting the production of vocal calls, which may represent an early scaffolding for reference signals, such as alarm calls depicting specific events (Fischer & Price, 2017; Froesel et al., 2022). An interesting instance is that of vervet monkeys, which have different calls depicting specific predators. While the structure of the calls is innate, their predator references are learned through social experience (Seyfarth et al., 1980). It remains to be determined whether the STS mediates the development of such referential communication as we would suggest.

In humans, there is considerable evidence that the semantic network partly relies on the activation of the STS, ATL, and MTG (Binder et al., 2009; Bryant & Preuss, 2018; Hodgson et al., 2023; Lambon Ralph, 2014; Lambon Ralph et al., 2017; for a complementary review on the evolution of STS and MTG for semantic, see Petrides, 2023). Furthermore, the ventral language stream is known to project to the pars triangularis (anterior Broca's area), receiving auditory and visual input via the extreme capsule, which participates in selecting lexico-semantic information in the human left hemisphere (Kostopoulos & Petrides, 2016). Rauschecker and Scott (2009) suggested that the sensorimotor role of the action stream in the visual system could be extended to the dorsal stream of the auditory cortex performing "auditory-motor transformations in verbal working memory tasks that involve articulatory representations" (p. 721). They argued that these transformations may be based on a multisensory reference frame. Therefore, we consider that in human brain evolution, the STS and neighboring sensorimotor and social brain areas were pivotal for the development of word-related concepts depicting objects, events, or names, which eventually expanded the semantic network into widespread representations across the cerebral cortex (Shahdloo et al., 2022; Small et al., 1995). It was recently reported that STS exhibits effective connectivity to regions of the inferior parietal lobe such as areas TPOJ1, STV, PSL and PGi (depicted in Figure 2), "which are language-related regions involved in semantic representations about objects, faces, and so on using multimodal information, and which then connect to Broca's area, especially to area 45" (Rauschecker & Afsahi, 2023, p. 1888). Interestingly, STS also shows effective connectivity to areas TGv and TGd in the ATL (Rolls et al., 2023).

Transmodal:  
Modality independent.

An influential theory in the field of semantic cognition recognizes the ATL as a modality invariant hub performing transmodal computations that are meant to control the use of semantic multisensory knowledge in a context-dependent manner (Lambon Ralph, 2014; Lambon Ralph et al., 2017). As mentioned earlier, both visual and auditory ventral streams seem to be organized along a posteroanterior gradient with increasingly complex processing of sensory inputs in anterior parts of the temporal lobe. According to these anatomo-functional schemes, supported by previous evidence reviewed in this perspective article, we do not

exclude the possibility of a similar organization for the semantic network. A posteroanterior gradient may exist in the human temporal lobe, from low-level sensorimotor integration of multimodal social signals (e.g., not only biological motion of bodies and faces but also voice processing) in posterior regions of the STS and MTG to higher-order, transmodal semantic control for time and context-appropriate use of multimodal experiential knowledge in anterior portions of the temporal lobe, especially ATL. For instance, if you order “some coffee” at a restaurant, it is very unlikely that the waiter will bring you a cup of coffee beans (which would be semantically correct but pragmatically clumsy). However, suppose you are a barista opening the coffee machine and you ask your colleague for some coffee. Hopefully, in that case, he will bring you some coffee beans to refill the machine bean container. According to the transmodal hub hypothesis, the ATL would supervise the pragmatic decision that better suits the context or task.

Despite promising progress in addressing some knowledge gaps, research on multimodal primate communication remains relatively uncommon. Some authors believe that theories of language evolution are unlikely to advance unless the field of primate communication research recognizes and addresses these knowledge gaps (Liebal et al., 2022). Theories suggesting that human language may have a multimodal origin are beginning to receive more support (Fröhlich et al., 2019; Wacewicz & Zywickzynski, 2017). This could introduce a theoretical shift and potentially promote the development of methodological approaches to empirically assess the multimodality of primate communication. In this perspective, we propose a new research agenda, focusing on the lateralization of a third lateral pathway running along the STS and TPJ, particularly concerning its functions in the language-dominant left hemisphere. Furthermore, comparative studies about this region may provide important insights into the evolutionary origin of our most cherished capacity, which is human language.

## FUNDING INFORMATION

Francisco Aboitiz, Agencia Nacional de Investigación y Desarrollo (<https://dx.doi.org/10.13039/501100020884>), Award ID: Fondecyt Regular 1210659. Maëva Michon, Agencia Nacional de Investigación y Desarrollo (<https://dx.doi.org/10.13039/501100020884>), Award ID: Fondecyt Postdoctoral 3201057.

## AUTHOR CONTRIBUTIONS

**Francisco Aboitiz:** Conceptualization: Equal; Funding acquisition: Equal; Investigation: Equal; Writing – original draft: Equal; Writing – review & editing: Equal. **Maëva Michon:** Conceptualization: Equal; Funding acquisition: Equal; Investigation: Equal; Writing – original draft: Equal; Writing – review & editing: Equal.

## REFERENCES

- Aboitiz, F. (2018a). A brain for speech. Evolutionary continuity in primate and human auditory-vocal processing. *Frontiers in Neuroscience*, 12, Article 174. <https://doi.org/10.3389/fnins.2018.00174>, PubMed: 29636657
- Aboitiz, F. (2018b). Voice, gesture and working memory in the emergence of speech. *Interaction Studies*, 19(1–2), 70–85. <https://doi.org/10.1075/is.17032.abo>
- Aboitiz, F., & García, R. (1997). The evolutionary origin of the language areas in the human brain. A neuroanatomical perspective. *Brain Research Reviews*, 25(3), 381–396. [https://doi.org/10.1016/S0165-0173\(97\)00053-2](https://doi.org/10.1016/S0165-0173(97)00053-2), PubMed: 9495565
- Agus, T. R., Paquette, S., Suied, C., Pressnitzer, D., & Belin, P. (2017). Voice selectivity in the temporal voice area despite matched low-level acoustic cues. *Scientific Reports*, 7(1), Article 11526. <https://doi.org/10.1038/s41598-017-11684-1>, PubMed: 28912437
- Arnott, S. R., Binns, M. A., Grady, C. L., & Alain, C. (2004). Assessing the auditory dual-pathway model in humans. *NeuroImage*,

- 22(1), 401–408. <https://doi.org/10.1016/j.neuroimage.2004.01.014>, PubMed: 15110033
- Bean, N. L., Smyre, S. A., Stein, B. E., & Rowland, B. A. (2023). Noise-rearing precludes the behavioral benefits of multisensory integration. *Cerebral Cortex*, 33(4), 948–958. <https://doi.org/10.1093/cercor/bhac113>, PubMed: 35332919
- Becker, Y., & Meguerditchian, A. (2022). Structural brain asymmetries for language: A comparative approach across primates. *Symmetry*, 14(5), Article 876. <https://doi.org/10.3390/sym14050876>
- Belin, P., Trapeau, R., & Obliger-Debouche, M. (2023). A small, but vocal, brain. *Cell Reports*, 42(6), Article 112651. <https://doi.org/10.1016/j.celrep.2023.112651>, PubMed: 37314925
- Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, 403(6767), 309–312. <https://doi.org/10.1038/35002078>, PubMed: 10659849
- Bernstein, L. E., Jiang, J., Pantazis, D., Lu, Z.-L., & Joshi, A. (2011). Visual phonetic processing localized using speech and non-speech face gestures in video and point-light displays. *Human Brain Mapping*, 32(10), 1660–1676. <https://doi.org/10.1002/hbm.21139>, PubMed: 20853377
- Bernstein, L. E., & Liebenthal, E. (2014). Neural pathways for visual speech perception. *Frontiers in Neuroscience*, 8, Article 386. <https://doi.org/10.3389/fnins.2014.00386>, PubMed: 25520611
- Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences*, 15(11), 527–536. <https://doi.org/10.1016/j.tics.2011.10.001>, PubMed: 22001867
- Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where Is the Semantic System? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, 19(12), 2767–2796. <https://doi.org/10.1093/cercor/bhp055>, PubMed: 19329570
- Bodin, C., Takerkart, S., Belin, P., & Coulon, O. (2018). Anatomofunctional correspondence in the superior temporal sulcus. *Brain Structure and Function*, 223(1), 221–232. <https://doi.org/10.1007/s00429-017-1483-2>, PubMed: 28756487
- Bodin, C., Trapeau, R., Nazarian, B., Sein, J., Degiovanni, X., Baurberg, J., Rapha, E., Renaud, L., Giordano, B. L., & Belin, P. (2021). Functionally homologous representation of vocalizations in the auditory cortex of humans and macaques. *Current Biology*, 31(21), 4839–4844. <https://doi.org/10.1016/j.cub.2021.08.043>, PubMed: 34506729
- Borghesani, A. M., Mazzuca, C., Gervasi, A. M., Mannella, F., & Tummolini, L. (2024). Grounded cognition can be multimodal all the way down. *Language, Cognition and Neuroscience*, 39(7), 838–842. <https://doi.org/10.1080/23273798.2023.2210238>
- Bruce, C., Desimone, R., & Gross, C. G. (1981). Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *Journal of Neurophysiology*, 46(2), 369–384. <https://doi.org/10.1152/jn.1981.46.2.369>, PubMed: 6267219
- Bryant, K. L., & Preuss, T. M. (2018). A comparative perspective on the human temporal lobe. In E. Bruner, N. Ogihara, & H. C. Tanabe (Eds.), *Digital endocasts: From skulls to brains* (pp. 239–258). Springer. [https://doi.org/10.1007/978-4-431-56582-6\\_16](https://doi.org/10.1007/978-4-431-56582-6_16)
- Calzavarini, F. (2024). Rethinking modality-specificity in the cognitive neuroscience of concrete word meaning: A position paper. *Language, Cognition and Neuroscience*, 39(7), 815–837. <https://doi.org/10.1080/23273798.2023.2173789>
- Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, 3(3), 201–215. <https://doi.org/10.1038/nrn755>, PubMed: 11994752
- Damera, S. R., Chang, L., Nikolov, P. P., Mattei, J. A., Banerjee, S., Glezer, L. S., Cox, P. H., Jiang, X., Rauschecker, J. P., & Riesenhuber, M. (2023). Evidence for a spoken word lexicon in the auditory ventral stream. *Neurobiology of Language*, 4(3), 420–434. [https://doi.org/10.1162/nol\\_a\\_00108](https://doi.org/10.1162/nol_a_00108), PubMed: 37588129
- DeWitt, I., & Rauschecker, J. P. (2012). Phoneme and word recognition in the auditory ventral stream. *Proceedings of the National Academy of Sciences*, 109(8), E505–E514. <https://doi.org/10.1073/pnas.1113427109>, PubMed: 22308358
- Dove, G. O. (2023a). Rethinking the role of language in embodied cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 378(1870), Article 20210375. <https://doi.org/10.1098/rstb.2021.0375>, PubMed: 36571130
- Dove, G. O. (2023b). Concepts require flexible grounding. *Brain and Language*, 245, Article 105322. <https://doi.org/10.1016/j.bandl.2023.105322>, PubMed: 37713771
- Dove, G. O. (2024). Grounding requires multimodal and multilevel representations. *Language, Cognition and Neuroscience*, 39(7), 843–846. <https://doi.org/10.1080/23273798.2023.2247501>
- Dubois, J., Benders, M., Lazeyras, F., Borradori-Tolsa, C., Leuchter, R. H.-V., Mangin, J. F., & Hüppi, P. S. (2010). Structural asymmetries of perisylvian regions in the preterm newborn. *NeuroImage*, 52(1), 32–42. <https://doi.org/10.1016/j.neuroimage.2010.03.054>, PubMed: 20362679
- Fernandino, L., Binder, J. R., Desai, R. H., Pendl, S. L., Humphries, C. J., Gross, W. L., Conant, L. L., & Seidenberg, M. S. (2016). Concept representation reflects multimodal abstraction: A framework for embodied semantics. *Cerebral Cortex*, 26(5), 2018–2034. <https://doi.org/10.1093/cercor/bhv020>, PubMed: 25750259
- Fernandino, L., & Conant, L. L. (2023). The primacy of experience in language processing: Semantic priming is driven primarily by experiential similarity. *bioRxiv*. <https://doi.org/10.1101/2023.03.21.533703>, PubMed: 36993310
- Fernandino, L., Tong, J.-Q., Conant, L. L., Humphries, C. J., & Binder, J. R. (2022). Decoding the information structure underlying the neural representation of concepts. *Proceedings of the National Academy of Sciences*, 119(6), Article e2108091119. <https://doi.org/10.1073/pnas.2108091119>, PubMed: 35115397
- Fischer, J., & Price, T. (2017). Meaning, intention, and inference in primate vocal communication. *Neuroscience & Biobehavioral Reviews*, 82, 22–31. <https://doi.org/10.1016/j.neubiorev.2016.10.014>, PubMed: 27773691
- Froesel, M., Gacoin, M., Clavagnier, S., Hauser, M., Goudard, Q., & Ben Hamed, S. (2022). Socially meaningful visual context either enhances or inhibits vocalisation processing in the macaque brain. *Nature Communications*, 13(1), Article 4886. <https://doi.org/10.1038/s41467-022-32512-9>, PubMed: 35985995
- Fröhlich, M., Sievers, C., Townsend, S. W., Gruber, T., & van Schaik, C. P. (2019). Multimodal communication and language origins: Integrating gestures and vocalizations. *Biological Reviews*, 94(5), 1809–1829. <https://doi.org/10.1111/brv.12535>, PubMed: 31250542
- Gandolfo, M., Abassi, E., Balgova, E., Downing, P. E., Papeo, L., & Koldewyn, K. (2024). Converging evidence that left extrastriate body area supports visual sensitivity to social interactions. *Current Biology*, 34(2), 343–351. <https://doi.org/10.1016/j.cub.2023.12.009>, PubMed: 38181794
- Ghazanfar, A. A., Chandrasekaran, C., & Logothetis, N. K. (2008). Interactions between the superior temporal sulcus and auditory cortex mediate dynamic face/voice integration in rhesus

- monkeys. *Journal of Neuroscience*, 28(17), 4457–4469. <https://doi.org/10.1523/JNEUROSCI.0541-08.2008>, PubMed: 18434524
- Ghazanfar, A. A., Maier, J. X., Hoffman, K. L., & Logothetis, N. K. (2005). Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *Journal of Neuroscience*, 25(20), 5004–5012. <https://doi.org/10.1523/JNEUROSCI.0799-05.2005>, PubMed: 15901781
- Gomez-Marin, A., & Ghazanfar, A. A. (2019). The life of behavior. *Neuron*, 104(1), 25–36. <https://doi.org/10.1016/j.neuron.2019.09.017>, PubMed: 31600513
- Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1), 20–25. [https://doi.org/10.1016/0166-2236\(92\)90344-8](https://doi.org/10.1016/0166-2236(92)90344-8), PubMed: 1374953
- Harpaintner, M., Sim, E.-J., Trumpp, N. M., Ulrich, M., & Kiefer, M. (2020). The grounding of abstract concepts in the motor and visual system: An fMRI study. *Cortex*, 124, 1–22. <https://doi.org/10.1016/j.cortex.2019.10.014>, PubMed: 31821905
- Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5), 393–402. <https://doi.org/10.1038/nrn2113>, PubMed: 17431404
- Hodgson, V. J., Lambon Ralph, M. A., & Jackson, R. L. (2023). The cross-domain functional organization of posterior lateral temporal cortex: Insights from ALE meta-analyses of 7 cognitive domains spanning 12,000 participants. *Cerebral Cortex*, 33(8), 4990–5006. <https://doi.org/10.1093/cercor/bhac394>, PubMed: 36269034
- Hopkins, W. D., Coulon, O., Meguerditchian, A., Staes, N., Sherwood, C. C., Schapiro, S. J., Mangin, J.-F., & Bradley, B. (2023). Genetic determinants of individual variation in the superior temporal sulcus of chimpanzees (*Pan troglodytes*). *Cerebral Cortex*, 33(5), 1925–1940. <https://doi.org/10.1093/cercor/bhac183>, PubMed: 35697647
- Ibáñez, A., Kühne, K., Miklashevsky, A., Monaco, E., Muraki, E., Ranzini, M., Speed, L. J., & Tuena, C. (2023). Ecological meanings: A consensus paper on individual differences and contextual influences in embodied language. *Journal of Cognition*, 6(1), Article 59. <https://doi.org/10.5334/joc.228>, PubMed: 37841670
- Jafari, A., Dureux, A., Zanini, A., Menon, R. S., Gilbert, K. M., & Everling, S. (2023). A vocalization-processing network in marmosets. *Cell Reports*, 42(5), Article 112526. <https://doi.org/10.1016/j.celrep.2023.112526>, PubMed: 37195863
- Jeschke, L., Mathias, B., & von Kriegstein, K. (2023). Inhibitory TMS over visual area V5/MT disrupts visual speech recognition. *Journal of Neuroscience*, 43(45), 7690–7699. <https://doi.org/10.1523/JNEUROSCI.0975-23.2023>, PubMed: 37848284
- Kaas, J. H., & Hackett, T. A. (1999). 'What' and 'where' processing in auditory cortex. *Nature Neuroscience*, 2(12), 1045–1047. <https://doi.org/10.1038/15967>, PubMed: 10570476
- Kausel, L., Michon, M., Soto-Icaza, P., & Aboitiz, F. (2024). A multimodal interface for speech perception: The role of the left superior temporal sulcus in social cognition and autism. *Cerebral Cortex*, 34(13), 84–93. <https://doi.org/10.1093/cercor/bhae066>, PubMed: 38696598
- Kewenig, V. N., Vigliocco, G., & Skipper, J. I. (2024). When abstract becomes concrete, naturalistic encoding of concepts in the brain. *eLife*, 13, Article RP91522. <https://doi.org/10.7554/eLife.91522>, PubMed: 39636743
- Khandhadia, A. P., Murphy, A. P., Romanski, L. M., Bizley, J. K., & Leopold, D. A. (2021). Audiovisual integration in macaque face patch neurons. *Current Biology*, 31(9), 1826–1835. <https://doi.org/10.1016/j.cub.2021.01.102>, PubMed: 33636119
- Kostopoulos, P., & Petrides, M. (2016). Selective memory retrieval of auditory what and auditory where involves the ventrolateral prefrontal cortex. *Proceedings of the National Academy of Sciences*, 113(7), 1919–1924. <https://doi.org/10.1073/pnas.1520432113>, PubMed: 26831102
- Kuhnke, P., Beaupain, M. C., Arola, J., Kiefer, M., & Hartwigsen, G. (2023). Meta-analytic evidence for a novel hierarchical model of conceptual processing. *Neuroscience & Biobehavioral Reviews*, 144, Article 104994. <https://doi.org/10.1016/j.neubiorev.2022.104994>, PubMed: 36509206
- Kuhnke, P., Kiefer, M., & Hartwigsen, G. (2020). Task-dependent recruitment of modality-specific and multimodal regions during conceptual processing. *Cerebral Cortex*, 30(7), 3938–3959. <https://doi.org/10.1093/cercor/bhaa010>, PubMed: 32219378
- Kuhnke, P., Kiefer, M., & Hartwigsen, G. (2023). Conceptual representations in the default, control and attention networks are task-dependent and cross-modal. *Brain and Language*, 244, Article 105313. <https://doi.org/10.1016/j.bandl.2023.105313>, PubMed: 37595340
- Lambon Ralph, M. A. (2014). Neurocognitive insights on conceptual knowledge and its breakdown. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1634), Article 20120392. <https://doi.org/10.1098/rstb.2012.0392>, PubMed: 24324236
- Lambon Ralph, M. A., Jeffries, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, 18(1), 42–55. <https://doi.org/10.1038/nrn.2016.150>, PubMed: 27881854
- Landsiedel, J., & Koldewyn, K. (2023). Auditory dyadic interactions through the "eye" of the social brain: How visual is the posterior STS interaction region? *Imaging Neuroscience*, 1, 1–20. [https://doi.org/10.1162/imag\\_a\\_00003](https://doi.org/10.1162/imag_a_00003), PubMed: 37719835
- Leaver, A. M., & Rauschecker, J. P. (2010). Cortical representation of natural complex sounds: Effects of acoustic features and auditory object category. *Journal of Neuroscience*, 30(22), 7604–7612. <https://doi.org/10.1523/JNEUROSCI.0296-10.2010>, PubMed: 20519535
- Leroy, F., Cai, Q., Bogart, S. L., Dubois, J., Coulon, O., Monzalvo, K., Fischer, C., Glasel, H., Van der Haegen, L., Bénézit, A., Lin, C.-P., Kennedy, D. N., Ihara, A. S., Hertz-Pannier, L., Moutard, M.-L., Poupon, C., Brysbaert, M., Roberts, N., Hopkins, W. D., ... Dehaene-Lambertz, G. (2015). New human-specific brain landmark: The depth asymmetry of superior temporal sulcus. *Proceedings of the National Academy of Sciences*, 112(4), 1208–1213. <https://doi.org/10.1073/pnas.1412389112>, PubMed: 25583500
- Liebal, K., Slocombe, K. E., & Waller, B. M. (2022). The language void 10 years on: Multimodal primate communication research is still uncommon. *Ethology Ecology & Evolution*, 34(3), 274–287. <https://doi.org/10.1080/03949370.2021.2015453>
- Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020). The Lancaster Sensorimotor Norms: Multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, 52(3), 1271–1291. <https://doi.org/10.3758/s13428-019-01316-z>, PubMed: 31832879
- McMahon, E., Bonner, M. F., & Isik, L. (2023). Hierarchical organization of social action features along the lateral visual pathway. *Current Biology*, 33(23), 5035–5047. <https://doi.org/10.1016/j.cub.2023.10.015>, PubMed: 37918399
- McMahon, E., & Isik, L. (2023). Seeing social interactions. *Trends in Cognitive Sciences*, 27(12), 1165–1179. <https://doi.org/10.1016/j.tics.2023.09.001>, PubMed: 37805385

- McMahon, E., & Isik, L. (2024). Abstract social interaction representations along the lateral pathway. *Trends in Cognitive Sciences*, 28(5), 392–393. <https://doi.org/10.1016/j.tics.2024.03.007>, PubMed: 38632007
- Meguerditchian, A. (2022). On the gestural origins of language: What baboons' gestures and brain have told us after 15 years of research. *Ethology Ecology & Evolution*, 34(3), 288–302. <https://doi.org/10.1080/03949370.2022.2044388>
- Meguerditchian, A., Marie, D., Margiotoudi, K., Roth, M., Nazarian, B., Anton, J.-L., & Claidière, N. (2021). Baboons (*Papio anubis*) living in larger social groups have bigger brains. *Evolution and Human Behavior*, 42(1), 30–34. <https://doi.org/10.1016/j.evolhumbehav.2020.06.010>
- Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, 343(6174), 1006–1010. <https://doi.org/10.1126/science.1245994>, PubMed: 24482117
- Michon, M., Billeke, P., Galgani, G., & Aboitiz, F. (2024). Sub-visemic discrimination and the effect of visual resemblance on silent lip-reading. *Language, Cognition and Neuroscience*, 39(5), 677–685. <https://doi.org/10.1080/23273798.2024.2353147>
- Michon, M., Boncompte, G., & López, V. (2020). Electrophysiological dynamics of visual speech processing and the role of orofacial effectors for cross-modal predictions. *Frontiers in Human Neuroscience*, 14, Article 538619. <https://doi.org/10.3389/fnhum.2020.538619>, PubMed: 33192386
- Michon, M., López, V., & Aboitiz, F. (2019). Origin and evolution of human speech: Emergence from a trimodal auditory, visual and vocal network. In M. A. Hofman (Ed.), *Evolution of the human brain: From matter to mind* (pp. 345–371). Elsevier. <https://doi.org/10.1016/bs.pbr.2019.01.005>, PubMed: 31703907
- Michon, M., Zamorano-Abramson, J., & Aboitiz, F. (2022). Faces and voices processing in human and primate brains: Rhythmic and multimodal mechanisms underlying the evolution and development of speech. *Frontiers in Psychology*, 13, Article 829083. <https://doi.org/10.3389/fpsyg.2022.829083>, PubMed: 35432052
- Miklashevsky, A. (2018). Perceptual experience norms for 506 Russian nouns: Modality rating, spatial localization, manipulability, imageability and other variables. *Journal of Psycholinguistic Research*, 47(3), 641–661. <https://doi.org/10.1007/s10936-017-9548-1>, PubMed: 29282595
- Mishkin, M., & Ungerleider, L. G. (1982). Contribution of striate inputs to the visuospatial functions of parieto-preoccipital cortex in monkeys. *Behavioural Brain Research*, 6(1), 57–77. [https://doi.org/10.1016/0166-4328\(82\)90081-X](https://doi.org/10.1016/0166-4328(82)90081-X), PubMed: 7126325
- Newell, F. N., McKenna, E., Seveso, M. A., Devine, I., Alahmad, F., Hirst, R. J., & O'Dowd, A. (2023). Multisensory perception constrains the formation of object categories: A review of evidence from sensory-driven and predictive processes on categorical decisions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 378(1886), Article 20220342. <https://doi.org/10.1098/rstb.2022.0342>, PubMed: 37545304
- Norman-Haignere, S., Kanwisher, N. G., & McDermott, J. H. (2015). Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *Neuron*, 88(6), 1281–1296. <https://doi.org/10.1016/j.neuron.2015.11.035>, PubMed: 26687225
- Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active inference: The free energy principle in mind, brain, and behavior*. MIT Press. <https://doi.org/10.7551/mitpress/12441.001.0001>
- Patel, G. H., Sestieri, C., & Corbetta, M. (2019). The evolution of the temporoparietal junction and posterior superior temporal sulcus. *Cortex*, 118, 38–50. <https://doi.org/10.1016/j.cortex.2019.01.026>, PubMed: 30808550
- Pernet, C. R., McAleer, P., Latinus, M., Gorgolewski, K. J., Charest, I., Bestelmeyer, P. E. G., Watson, R. H., Fleming, D., Crabbe, F., Valdes-Sosa, M., & Belin, P. (2015). The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices. *NeuroImage*, 119, 164–174. <https://doi.org/10.1016/j.neuroimage.2015.06.050>, PubMed: 26116964
- Perrotin, C., Kayser, C., Abel, T. J., Logothetis, N. K., & Petkov, C. I. (2015). Who is that? Brain networks and mechanisms for identifying individuals. *Trends in Cognitive Sciences*, 19(12), 783–796. <https://doi.org/10.1016/j.tics.2015.09.002>, PubMed: 26454482
- Perrotin, C., Kayser, C., Logothetis, N. K., & Petkov, C. I. (2014). Auditory and visual modulation of temporal lobe neurons in voice-sensitive and association cortices. *Journal of Neuroscience*, 34(7), 2524–2537. <https://doi.org/10.1523/JNEUROSCI.2805-13.2014>, PubMed: 24523543
- Petrides, M. (2014). *Neuroanatomy of language regions of the human brain*. Academic Press. <https://doi.org/10.1016/C2011-0-07354-4>
- Petrides, M. (2023). On the evolution of polysensory superior temporal sulcus and middle temporal gyrus: A key component of the semantic system in the human brain. *Journal of Comparative Neurology*, 531(18), 1987–1995. <https://doi.org/10.1002/cne.25521>, PubMed: 37434287
- Pezzulo, G., Parr, T., Cisek, P., Clark, A., & Friston, K. (2024). Generating meaning: Active inference and the scope and limits of passive AI. *Trends in Cognitive Sciences*, 28(2), 97–112. <https://doi.org/10.1016/j.tics.2023.10.002>, PubMed: 37973519
- Pitcher, D. (2021). Characterizing the third visual pathway for social perception. *Trends in Cognitive Sciences*, 25(7), 550–551. <https://doi.org/10.1016/j.tics.2021.04.008>, PubMed: 34024729
- Pitcher, D. (2023). Visual neuroscience: A specialised neural pathway for social perception. *Current Biology*, 33(23), R1222–R1224. <https://doi.org/10.1016/j.cub.2023.10.020>, PubMed: 38052168
- Pitcher, D., & Ungerleider, L. G. (2021). Evidence for a third visual pathway specialized for social perception. *Trends in Cognitive Sciences*, 25(2), 100–110. <https://doi.org/10.1016/j.tics.2020.11.006>, PubMed: 33334693
- Pouw, W., Proksch, S., Drijvers, L., Gamba, M., Holler, J., Kello, C., Schaefer, R. S., & Wiggins, G. A. (2021). Multilevel rhythms in multimodal communication. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376(1835), Article 20200334. <https://doi.org/10.1098/rstb.2020.0334>, PubMed: 34420378
- Pulvermüller, F. (2013). How neurons make meaning: Brain mechanisms for embodied and abstract-symbolic semantics. *Trends in Cognitive Sciences*, 17(9), 458–470. <https://doi.org/10.1016/j.tics.2013.06.004>, PubMed: 23932069
- Pulvermüller, F., & Fadiga, L. (2010). Active perception: Sensorimotor circuits as a cortical basis for language. *Nature Reviews Neuroscience*, 11(5), 351–360. <https://doi.org/10.1038/nrn2811>, PubMed: 20383203
- Ramstead, M. J., Kirchhoff, M. D., & Friston, K. J. (2020). A tale of two densities: Active inference is enactive inference. *Adaptive Behavior*, 28(4), 225–239. <https://doi.org/10.1177/1059712319862774>, PubMed: 32831534
- Rauschecker, J. P., & Afsahi, R. K. (2023). Anatomy of the auditory cortex then and now. *Journal of Comparative Neurology*, 531(18), 1883–1892. <https://doi.org/10.1002/cne.25560>, PubMed: 38010215

- Rauschecker, J. P., & Scott, S. K. (2009). Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing. *Nature Neuroscience*, 12(6), 718–724. <https://doi.org/10.1038/nn.2331>, PubMed: 19471271
- Rauschecker, J. P., & Tian, B. (2000). Mechanisms and streams for processing of “what” and “where” in auditory cortex. *Proceedings of the National Academy of Sciences*, 97(22), 11800–11806. <https://doi.org/10.1073/pnas.97.22.11800>, PubMed: 11050212
- Rauschecker, J. P., Tian, B., & Hauser, M. (1995). Processing of complex sounds in the macaque nonprimary auditory cortex. *Science*, 268(5207), 111–114. <https://doi.org/10.1126/science.7701330>, PubMed: 7701330
- Redcay, E. (2008). The superior temporal sulcus performs a common function for social and speech perception: Implications for the emergence of autism. *Neuroscience & Biobehavioral Reviews*, 32(1), 123–142. <https://doi.org/10.1016/j.neubiorev.2007.06.004>, PubMed: 17706781
- Rennig, J., & Beauchamp, M. S. (2018). Free viewing of talking faces reveals mouth and eye preferring regions of the human superior temporal sulcus. *NeuroImage*, 183, 25–36. <https://doi.org/10.1016/j.neuroimage.2018.08.008>, PubMed: 30092347
- Rennig, J., & Beauchamp, M. S. (2022). Intelligibility of audiovisual sentences drives multivoxel response patterns in human superior temporal cortex. *NeuroImage*, 247, Article 118796. <https://doi.org/10.1016/j.neuroimage.2021.118796>, PubMed: 34906712
- Rilling, J. K., Glasser, M. F., Preuss, T. M., Ma, X., Zhao, T., Hu, X., & Behrens, T. E. J. (2008). The evolution of the arcuate fasciculus revealed with comparative DTI. *Nature Neuroscience*, 11(4), 426–428. <https://doi.org/10.1038/nn2072>, PubMed: 18344993
- Rolls, E. T., Rauschecker, J. P., Deco, G., Huang, C.-C., & Feng, J. (2023). Auditory cortical connectivity in humans. *Cerebral Cortex*, 33(10), 6207–6227. <https://doi.org/10.1093/cercor/bhac496>, PubMed: 36573464
- Roumazeilles, L., Eichert, N., Bryant, K. L., Folloni, D., Sallet, J., Vijayakumar, S., Foxley, S., Tendler, B. C., Jbabdi, S., Reveley, C., Verhagen, L., Dershowitz, L. B., Guthrie, M., Flach, E., Miller, K. L., & Mars, R. B. (2020). Longitudinal connections and the organization of the temporal cortex in macaques, great apes, and humans. *PLOS Biology*, 18(7), Article e3000810. <https://doi.org/10.1371/journal.pbio.3000810>, PubMed: 32735557
- Roumazeilles, L., Schurz, M., Lojkiewicz, M., Verhagen, L., Schüffelgen, U., Marche, K., Mahmoodi, A., Emberton, A., Simpson, K., Joly, O., Khamassi, M., Rushworth, M. F. S., Mars, R. B., & Sallet, J. (2021). Social prediction modulates activity of macaque superior temporal cortex. *Science Advances*, 7(38), Article eabh2392. <https://doi.org/10.1126/sciadv.abh2392>, PubMed: 34524842
- Sallet, J., Mars, R. B., Noonan, M. P., Andersson, J. L., O'Reilly, J. X., Jbabdi, S., Croxson, P. L., Jenkinson, M., Miller, K. L., & Rushworth, M. F. S. (2011). Social network size affects neural circuits in macaques. *Science*, 334(6056), 697–700. <https://doi.org/10.1126/science.1210027>, PubMed: 22053054
- Schroeder, C. E., Wilson, D. A., Radman, T., Scharfman, H., & Lakatos, P. (2010). Dynamics of active sensing and perceptual selection. *Current Opinion in Neurobiology*, 20(2), 172–176. <https://doi.org/10.1016/j.conb.2010.02.010>, PubMed: 20307966
- Scott, B. H., Mishkin, M., & Yin, P. (2012). Monkeys have a limited form of short-term memory in audition. *Proceedings of the National Academy of Sciences*, 109(30), 12237–12241. <https://doi.org/10.1073/pnas.1209685109>, PubMed: 22778411
- Seyfarth, R. M., Cheney, D. L., & Marler, P. (1980). Monkey responses to three different alarm calls: Evidence of predator classification and semantic communication. *Science*, 210(4471), 801–803. <https://doi.org/10.1126/science.7433999>, PubMed: 7433999
- Shahdloo, M., Çelik, E., Urgen, B. A., Gallant, J. L., & Çukur, T. (2022). Task-dependent warping of semantic representations during search for visual action categories. *Journal of Neuroscience*, 42(35), 6782–6799. <https://doi.org/10.1523/JNEUROSCI.1372-21.2022>, PubMed: 35863889
- Sierpowska, J., Bryant, K. L., Janssen, N., Blazquez Freches, G., Römkens, M., Mangnus, M., Mars, R. B., & Piai, V. (2022). Comparing human and chimpanzee temporal lobe neuroanatomy reveals modifications to human language hubs beyond the frontotemporal arcuate fasciculus. *Proceedings of the National Academy of Sciences*, 119(28), Article e2118295119. <https://doi.org/10.1073/pnas.2118295119>, PubMed: 35787056
- Skeide, M. A., & Friederici, A. D. (2016). The ontogeny of the cortical language network. *Nature Reviews Neuroscience*, 17(5), 323–332. <https://doi.org/10.1038/nrn.2016.23>, PubMed: 27040907
- Small, S. L., Hart, J., Nguyen, T., & Gordon, B. (1995). Distributed representations of semantic knowledge in the brain. *Brain*, 118(2), 441–453. <https://doi.org/10.1093/brain/118.2.441>, PubMed: 7735885
- Smyre, S. A., Bean, N. L., Stein, B. E., & Rowland, B. A. (2024). The brain can develop conflicting multisensory principles to guide behavior. *Cerebral Cortex*, 34(6), Article bhae247. <https://doi.org/10.1093/cercor/bhae247>, PubMed: 38879756
- Speed, L. J., & Brysbaert, M. (2024). Ratings of valence, arousal, happiness, anger, fear, sadness, disgust, and surprise for 24,000 Dutch words. *Behavior Research Methods*, 56(5), 5023–5039. <https://doi.org/10.3758/s13428-023-02239-6>, PubMed: 37783901
- Speed, L. J., & Majid, A. (2017). Dutch modality exclusivity norms: Simulating perceptual modality in space. *Behavior Research Methods*, 49(6), 2204–2218. <https://doi.org/10.3758/s13428-017-0852-3>, PubMed: 28155185
- Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses* (p. xv, 211). MIT Press.
- Stein, B. E., Stanford, T. R., & Rowland, B. A. (2020). Multisensory integration and the society for neuroscience: Then and now. *Journal of Neuroscience*, 40(1), 3–11. <https://doi.org/10.1523/JNEUROSCI.0737-19.2019>, PubMed: 31676599
- Tian, B., Reser, D., Durham, A., Kustov, A., & Rauschecker, J. P. (2001). Functional specialization in rhesus monkey auditory cortex. *Science*, 292(5515), 290–293. <https://doi.org/10.1126/science.1058911>, PubMed: 11303104
- Tong, J., Binder, J. R., Humphries, C., Mazurchuk, S., Conant, L. L., & Fernandino, L. (2022). A distributed network for multimodal experiential representation of concepts. *Journal of Neuroscience*, 42(37), 7121–7130. <https://doi.org/10.1523/JNEUROSCI.1243-21.2022>, PubMed: 35940877
- Varela, F. J., Thompson, E., & Rosch, E. (2017). *The embodied mind: Cognitive science and human experience* (Revised ed.). MIT Press. <https://doi.org/10.7551/mitpress/9780262529365.001.0001>
- Vergallito, A., Petilli, M. A., & Marelli, M. (2020). Perceptual modality norms for 1,121 Italian words: A comparison with concreteness and imageability scores and an analysis of their impact in word processing tasks. *Behavior Research Methods*, 52(4), 1599–1616. <https://doi.org/10.3758/s13428-019-01337-8>, PubMed: 31950360
- Wacewicz, S., & Zywickzynski, P. (2017). The multimodal origins of linguistic communication. *Language & Communication*, 54, 1–8. <https://doi.org/10.1016/j.langcom.2016.10.001>

- Watson, R., Latinus, M., Charest, I., Crabbe, F., & Belin, P. (2014). People-selectivity, audiovisual integration and heteromodality in the superior temporal sulcus. *Cortex*, 50, 125–136. <https://doi.org/10.1016/j.cortex.2013.07.011>, PubMed: 23988132
- Weiner, K. S., & Grill-Spector, K. (2013). Neural representations of faces and limbs neighbor in human high-level visual cortex: Evidence for a new organization principle. *Psychological Research*, 77(1), 74–97. <https://doi.org/10.1007/s00426-011-0392-x>, PubMed: 22139022
- Wurm, M. F., & Caramazza, A. (2022). Two ‘what’ pathways for action and object recognition. *Trends in Cognitive Sciences*, 26(2), 103–116. <https://doi.org/10.1016/j.tics.2021.10.003>, PubMed: 34702661
- Zhong, Y., Wan, M., Ahrens, K., & Huang, C.-R. (2022). Sensorimotor norms for Chinese nouns and their relationship with orthographic and semantic variables. *Language, Cognition and Neuroscience*, 37(8), 1000–1022. <https://doi.org/10.1080/23273798.2022.2035416>
- Zhu, L. L., & Beauchamp, M. S. (2017). Mouth and voice: A relationship between visual and auditory preference in the human superior temporal sulcus. *Journal of Neuroscience*, 37(10), 2697–2708. <https://doi.org/10.1523/JNEUROSCI.2914-16.2017>, PubMed: 28179553