Downloaded from www.annualreviews.org. Stanford University Libraries (ar-224679) IP: 171.66.13.59 On: Fri, 25 Oct 2024 15:37:33

ANNUAL REVIEWS

Annual Review of Linguistics

Computational Architecture of Speech Comprehension in the Human Brain

Laura Gwilliams,[1,2,*] Ilina Bhaya-Grossman,[2,3,*] Yizhen Zhang,[2,*] Terri Scott,[2,4,*] Sarah Harper,[2,*] and Deborah Levy[2,5,*]

[1]Department of Psychology, Stanford University, Stanford, California, USA; email: laura.gwilliams@stanford.edu

[2]Department of Neurological Surgery, University of California, San Francisco, California, USA; email: ilina.bhaya-grossman@ucsf.edu, yizhen.zhang@ucsf.edu, sarah.ko.harper@gmail.com

[3]University of California, Berkeley–University of California, San Francisco Graduate Program in Bioengineering, Berkeley, California, USA

[4]Current affiliation: Department of Psychology, Northeastern University Oakland, Oakland, California, USA; email: te.scott@northeastern.edu

[5]Current affiliation: Princeton Writing Program, Princeton University, Princeton, New Jersey, USA; email: deborah.levy@princeton.edu

Annu. Rev. Linguist. 2025. 11:11.1–11.18

The *Annual Review of Linguistics* is online at linguistics.annualreviews.org

https://doi.org/10.1146/annurev-linguistics-031120-111245

Copyright © 2025 by the author(s).
All rights reserved

*All authors contributed equally to this article

Keywords

speech, computation, comprehension, processing, language, machine learning

Abstract

Understanding the computational algorithm that gives rise to human language is a shared endeavor among neuroscience, linguistics, and machine learning. We propose a conceptual framework for making measurable progress toward this goal by studying the subcomponents of the processing system: its underlying representations, operations, and information flow. We review evidence from neurophysiology, neuropsychology, linguistic theory, and computational modeling and suggest future directions to push the field forward in developing a precise characterization of spoken language understanding. Overall, we claim that representations of speech properties, and the operations that generate and manipulate those representations, exist within a highly parallel, highly redundant spatiotemporal regime.

Review in Advance. Changes may still occur before final publication.

Downloaded from www.annualreviews.org. Stanford University Libraries (ar-224679) IP: 171.66.13.59 On: Fri, 25 Oct 2024 15:37:33

INTRODUCTION

For a listener, understanding speech typically feels like a spontaneous, effortless process. The fluctuations in air pressure created by an interlocutor can instill colorful stories or mundane pleasantries without ever consciously commanding the mind to do so. This automaticity of understanding is in stark contrast to the computational complexity of transforming sounds into meaning.

There are many things that make speech comprehension a complicated task. First, there is nothing inherent in the sounds of language themselves, or the order in which they are articulated, that determines what meaning a person is trying to convey. This arbitrary mapping is demonstrated when conveying a similar concept (e.g., happiness) in different languages. Doing so requires employing different sounds in different systematic configurations (e.g., *felicidade* in Portuguese, *zoriona* in Basque, *Glück* in German). It is not the sounds that determine meaning but rather their learned and arbitrary associations (Holdcroft 1991).

Another significant challenge comes from the high variability in the acoustic realization of the same utterance, both within and across speakers. For instance, though the authors of this review all have similar-sized vocal tracts, the temporal-spectral analysis of each author's utterance of *happiness* looks very different; yet, all need to be mapped to a common conceptual representation (Klatt 1986, Liberman et al. 1952). This many-to-one mapping between acoustic realization and language unit requires abstraction over the sensory input in order to identify a linguistic sequence that is invariant to the specific acoustic realization.

Furthermore, though the speech signal is continuous, both in time and in modulation value, it must be transformed into discrete units, which connect to stored representations in the brain. There are no systematic silences between meaningful units in continuous speech, and so understanding where a relevant unit of language begins and ends, such that it can be scrutinized appropriately, is not trivial (Brent 1999).

How does the brain overcome these challenges and achieve speech comprehension? Given that there is no direct linear mapping between sound and meaning, the brain needs to abstract away from the sensory realization of the speech input in order to recognize the higher-order properties it contains. To do so, it must apply a series of nonlinear transformations on the acoustic signal to create a hierarchy of acoustic and linguistic representations that become less similar to the input, and more similar to the intended meaning, with each transformational step. In this review, we propose that these processes, which make up the computational architecture of speech comprehension, are best understood under a three-level framework: representations, operations, and information flow (see **Figure 1**). Studying the system relative to these fundamental components and how they interact provides a tractable analytical framework to gain understanding of the system as a whole and of how the system has evolved to solve the numerous challenges of speech comprehension.

REPRESENTATIONS, OPERATIONS, AND INFORMATION FLOW

What set of representations does the brain generate from the auditory signal to bridge from sound to meaning? Representations comprise properties of the speech signal, and properties of language, that the brain generates or retrieves during speech processing. Examples of representations could include a cochlear representation of the speech signal or something more abstract and linguistically motivated, such as syllables and morphemes (see **Figure 1**). Hypotheses of what those representations may be, as discussed below, have been generated based on linguistics, domain-general auditory neuroscience, and most recently computational approaches including deep language models and automatic speech recognition systems.

Review in Advance. Changes may still occur before final publication.

Downloaded from www.annualreviews.org. Stanford University Libraries (ar-224679) IP: 171.66.13.59 On: Fri, 25 Oct 2024 15:37:33

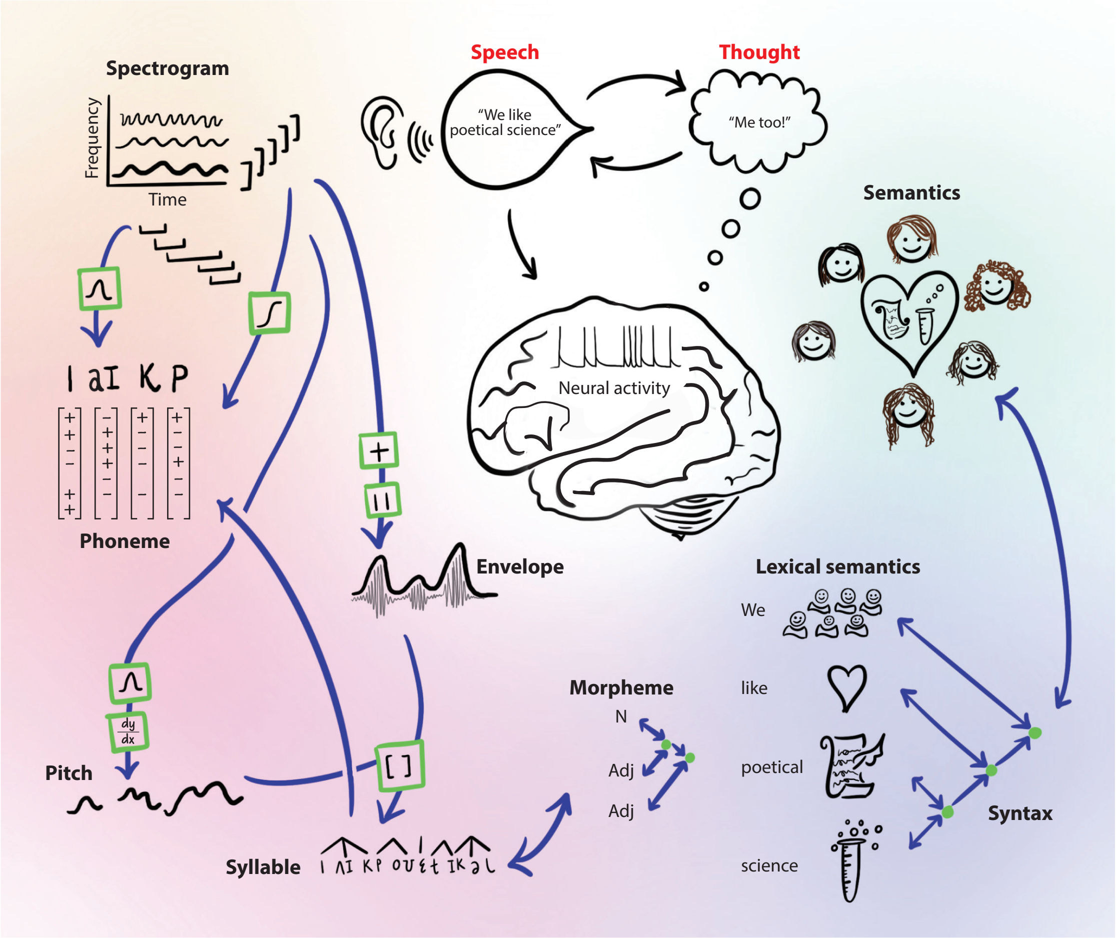

Figure 1

Schematic of the representations, operations, and information flow supporting comprehension of the spoken utterance *We like poetical science*. Candidate representations are shown in black, operations are shown in green boxes, and information flow is indicated by the blue arrows. Note that the information flow contains bidirectional connections, skip connections, and parallel processes.

What operations does the brain implement to generate and manipulate those representations? By operations we mean processes such as filtering the signal within a given frequency range, addition, multiplication, concatenation, or application of a nonlinear categorization. Each operation takes an input representation and performs a transformation to produce an output representation (see green boxes in **Figure 1**). By applying a large number of operations on the speech signal, whereby the output of one operation becomes the input to many others, the system is capable of generating very complex nonlinear properties of speech. Each operation takes a given amount of time to complete, contributing to overall processing time and cumulatively producing a reaction time that could be measured behaviorally on a given task.

Finally, it is crucial to determine the information flow of the system: In what order are representations generated, and in what order are operations performed? Even with the same set of representational bases and operations in place, depending on the dependency structure between them (e.g., representation A is a prerequisite for generating representation B), the overall output

Review in Advance. Changes may still occur before final publication.

Downloaded from www.annualreviews.org. Stanford University Libraries (ar-224679) IP: 171.66.13.59 On: Fri, 25 Oct 2024 15:37:33

of the system and the strategies it uses during processing could be drastically different. Representations can be passed between and within brain areas through the use of different forms of connection. These include feedforward connections, where information that has undergone fewer transformations (i.e., is relatively more "simple") is passed to neural populations that process relatively more "complex" information; feedback connections that pass complex information to neural populations that process simpler information; and recurrent connections that pass information at any level of complexity from one neural population back to the same neural population, thus allowing information to be maintained over time. Skip connections allow information to be passed between neural populations that are not immediately adjacent in the anatomical hierarchy. Connections are depicted in **Figure 1** as blue arrows.

Unlike representations and operations, which are defined over a single neural population at a given time, information flow is defined relative to how information moves across space and across time. This could be at the micro scale (e.g., within the processing circuit of a single cortical column), at the meso scale (e.g., information moving within a given gyrus or brain region), or at a more global scale (e.g., information passing from one brain region to another). Information flow can be studied on the basis of relative differences in the latency of neural responses, as time-locked to a given speech event, under the assumption that, as information passes across neural populations, it causes a cascade of responses that unfold in time.

Representations form the basis upon which to identify the operations and information flow of the system. Given that representations are the input and output of a given computation—one representation comes in and a different representation comes out—if the representation changes, or if the way it is encoded in neural activity (i.e., its coding scheme) changes, then a neural operation is at work. Similarly, the relative latency with which representations emerge in neural activity reflects how information is passing through the system to derive those representations.

Identifying the representations of the system is therefore the analytical bedrock upon which the other components of the system can be uncovered. Linguistic theory has proven to be an invaluable hypothesis generator for the types of representations that the brain may encode to process and understand language. Here, we review neurophysiological data from healthy adults, whereby modulation in neural activity during typical functioning is associated with the different hypothesized components of the speech input being processed. We also review evidence from adults with acquired language disorders, whereby lesions to certain brain areas can be associated with deficits in speech processing, thus providing causal links between the affected area and (*a*) the representations it houses, (*b*) the operations it likely performs, and (*c*) the dynamic topography of information flow. The common goal is to link the property of speech on the one hand (be it acoustic, linguistic, or statistical in nature) with the neural correlate on the other (be it strength of activity in a given area of the brain at a given latency or the pathology of, or damage to, neural tissue).

NEURAL SUBSTRATES OF THE COMPUTATIONAL ARCHITECTURE

Acoustic, Phonetic, and Syllabic Representations Are Polymorphic Across the Auditory Cortex

Linguistic theory tends to conceptualize acoustic, phonetic, and syllabic representations as separable from each other with each composed to form the next in a hierarchical manner. However, recent research shows a much more nuanced picture in which acoustic, phonetic, and syllabic representations exist in parallel and for long durations during speech processing. In this section we explore what is currently known about these sub- and suprasegmental features and their instantiation in the cortex; we emphasize the polymorphism of these representations—that is, the

Review in Advance. Changes may still occur before final publication.

Downloaded from www.annualreviews.org. Stanford University Libraries (ar-224679) IP: 171.66.13.59 On: Fri, 25 Oct 2024 15:37:33

parallel generation of distinct but highly redundant representations of the input that emphasize or minimize particular aspects of the signal, which can be used to optimize different downstream processes.

At early levels of speech perception (i.e., as sound enters the auditory cortex), representations are predominantly organized by frequency. As the sound signal reaches the auditory cortex, neural populations represent information about the frequency content of the sound and how it varies over time. This is inherited from the frequency-organized information derived at the cochlea in the inner ear and the subsequent representations of frequency in subcortical structures like the auditory thalamus. **Figure 1** illustrates this as a frequency-by-time representation of the input; spectral detail is analyzed across windows of time and across windows of frequency modulation (Whiteford et al. 2020). This spectrotemporal information has been shown to be encoded in the primary auditory cortex, where neurons respond to a narrow band of frequency information (Khalighinejad et al. 2021) with an average peak latency of ~40 ms (Simon et al. 2022).

This time–frequency representation of acoustic input is not specific to speech and is evoked in response to all acoustic inputs; however, some studies have found stronger responses to natural speech than to nonspeech stimuli (Belin & Zatorre 2000, Khalighinejad et al. 2021, Moerel et al. 2012). Furthermore, some studies have found that responses to vowels in the primary auditory cortex do not scale linearly with the acoustic input but rather exhibit categorical responses that align with perceptual vowel categories (Levy & Wilson 2020). This has also been demonstrated for nonspeech stimuli, where the participant learns new sound categories of complex sound ripples (Ley et al. 2012); such findings suggest that categorical responses likely arise to speech due to the learned categorical associations rather than because the stimulus is speech per se. Overall, representations of speech in the primary auditory cortex are predominantly frequency based, though they also exhibit nonlinear warping depending on learned categories (Coffey et al. 2016).

As we move away from the primary auditory cortex and into higher-order auditory regions such as the superior temporal gyrus (STG), representations become more multifaceted and complex. For speech inputs, this corresponds to receptive fields that contain multiple spectrotemporal peaks (Leonard et al. 2024) corresponding to the phonetic features of each individual speech sound. In contrast to the narrow-band frequency representations observed in the primary auditory cortex, neural populations in the STG—a nonprimary auditory region just lateral to the primary auditory cortex—exhibit receptive fields with complex modulation across both time and frequency (Hamilton et al. 2021). As a consequence of their complex spectrotemporal tuning, neural responses in the STG are stronger for complex sounds, including speech, than for narrow-frequency pure tones or noise (Zatorre et al. 1992), and thus they are ideally suited to discriminate speech sound sequences in terms of their phonetic content.

Importantly, the generation of these more complex acoustic representations does not entail deleting or discarding the lower-level narrow-band representations (Gwilliams et al. 2018); rather, the lower-level representations have been found to be encoded in parallel to more complex representations of the speech signal and encoded in neural activity for a long period of time.

The encoding of phonetic features in the STG has been shown both at the level of single neurons (Leonard et al. 2024) and at the level of populations of neurons as recorded from the cortical surface (Mesgarani et al. 2014, Oganian et al. 2023). Deriving features from the time–frequency representation requires integrating information over time and emphasizing certain components of the spectral content, as shown in **Figure 1**. By comparing the representations of these features in the STG across the micro and meso scales (single neurons versus tens of thousands), we can determine how the operations occurring at a single cortical column are integrated to give rise to representations at the surface level.

Review in Advance. Changes may
still occur before final publication.

Downloaded from www.annualreviews.org. Stanford University Libraries (ar-224679) IP: 171.66.13.59 On: Fri, 25 Oct 2024 15:37:33

Populations of neurons as measured using electrocorticography (ECoG) have been shown to respond to the phonetic features shared between phoneme categories rather than encoding phoneme categories per se. For example, responses at a given electrode will be shared for all phonemes that share a given feature (e.g., voiced, fricative, obstruent) (Mesgarani et al. 2014). The encoding of phonetic features and their associated spectrotemporal receptive fields are intermixed in a "salt and pepper" fashion (Mesgarani et al. 2014) with an overarching spatial organization of the posterior STG encoding phonetic features at onset and the anterior STG encoding phonetic features postonset (Hamilton et al. 2018).

When zooming in to the level of single neurons in the STG, phonetic features remain the appropriate representational format, with no evidence for phoneme category encoding (Leonard et al. 2024). When investigating single neuron encoding across different cortical layers, Leonard et al. (2024) observed a high degree of heterogeneity in the tuning of different neurons, whereby not all neurons were tuned to the same speech feature: Some neurons encoded phonetic features, whereas others encoded loudness, and others encoded the speaker's pitch of voice. Furthermore, this tuning was not intermixed as was observed across the surface, but rather, neurons that encoded similar speech properties localized to similar depths relative to the cortical surface. Overall, these results suggest that the fundamental speech feature representation in the STG is not only the phonetic feature but comprises other, highly redundant properties of speech as well. The fact that all of these features spatially co-localize to the same cortical column emphasizes that polymorphic representations not only are parallel in time but also are parallel in space, at the scale of microns.

These electrophysiology findings are corroborated by evidence from lesion studies. When individuals suffer left posterior STG damage poststroke or due to neurodegenerative disease, it can lead to deficits in phoneme discrimination (Blumstein et al. 1977, Johnson et al. 2020, Mesulam et al. 2019a, Robson et al. 2013). Furthermore, direct cortical stimulation to the left STG has been shown to impair the ability to discriminate between consonants (Boatman et al. 1997) and recognize distinct allophones as the same phoneme (Boatman 2004). This impaired phonemic processing may underlie deficits in single-word comprehension, as the ability to recognize a word often depends on perception of a single phoneme (e.g., *boat* /boʊt/ versus *bone* /boʊn/). A fascinating speech perception disorder, pure word deafness (PWD), also sheds light on these phonemic processes. PWD is characterized by profound difficulties in understanding speech, though other types of auditory perception and linguistic processing are relatively spared (Dumanch & Poling 2019, Poeppel 2001). Patients with PWD are aware of a sound when presented with speech, but it is not intelligible to them; the perception has been described as "jabber" and "muffled" (Buchman et al. 1986) or "meaningless. . .garbled sound" (Mendez & Rosenberg 1991) (see also Poeppel 2001). While PWD has commonly been believed to require bilateral lesions to the auditory cortex to result in the disorder, recent work suggests that damage to the left STG alone is sufficient to cause PWD-like symptoms (Casilio et al. 2024). Thus, neural populations in the left STG not only are active during phonetic processing but also are causally involved in its success.

Importantly, the feature representations in the STG have been shown to reflect not only the time window of the speech sound itself but also information that occurs before (Keshishian et al. 2020) and even after the speech sound (Gwilliams et al. 2018). This means that the representations in these neural populations are not strict static filters on the input but rather adapt in a state-dependent manner. This could manifest in terms of spatially intermixed phonetic representations interacting with one another (Bhaya-Grossman & Chang 2022, Yi et al. 2019) due to normalization for other speech properties like pitch range (Johnson & Sjerps 2021) or due to integration on information over an extended time window (Gwilliams et al. 2018). This is important because it shows that the encoding of speech features in the STG is not just a recapitulation of their complex spectrotemporal properties; rather, they are maintained in parallel with incoming speech input for

Review in Advance. Changes may still occur before final publication.

Downloaded from www.annualreviews.org. Stanford University Libraries (ar-224679) IP: 171.66.13.59 On: Fri, 25 Oct 2024 15:37:33

the purposes of extracting higher-order structures of speech across sequences of units and nested timescales (Gwilliams et al. 2022).

Another important cue in the speech signal represented in the primary cortex and STG is the speech envelope. The envelope is computed as a weighted average of the energy across different frequency bands in the input; spectral detail is removed, but critical temporal detail is retained (see **Figure 1**). The rhythmic structure of the speech envelope reflects the linguistic unit of the syllable. The syllable functions as an organizational unit with a critical role in grouping phonemes into sequences. Importantly, the syllable conveys meaning beyond its internal phonemic content (Blevins 1995), serving as a prosodic unit, relaying within-word stress information, and acting as a structural scaffold for its phonemic constituents. Time-resolved neural recording techniques, including electroencephalography, magnetoencephalography (MEG), and ECoG, have shown that there is a representation of the syllable dissociable from, and extracted in parallel to, phonetic content (Doelling et al. 2014, Hamilton et al. 2021, Luo & Poeppel 2007, Oganian & Chang 2019). The precise operation that derives syllabic information, and therefore the underlying format of the syllabic representation, remains a matter of heated debate—namely, between an operation that continuously tracks envelope phase information and one that tracks discrete and sparse syllabic events. It is likely that both representations exist in parallel in distinct physiological formats (Ray & Maunsell 2011).

The amplitude envelope is one of the strongest signals in the acoustic input, and it is encoded very strongly in neural activity in auditory cortices (Kubanek et al. 2013). Research suggests that altering the speech envelope of the input leads to issues in consonant recognition, vowel clarity, and sentence comprehension (Ahissar et al. 2001; Drullman et al. 1994a,b), and individuals can comprehend speech that maintains its temporal envelope despite significantly compromised frequency details (Shannon et al. 1995). Together, all of the above suggests that the speech envelope is an important auditory feature for speech comprehension.

Segmentation, Look-Up, and Composition of Morphemes Are Operations Underlying Word Processing

The features discussed so far are robustly encoded in the acoustic signal, and therefore simple linear or shallow nonlinear computations are sufficient to extract these features from the input. The arbitrary nature of sound–meaning mapping requires, however, that symbolic and abstract features are derived from the acoustic signal to make contact with the higher-order semantic and structural information it contains (Barsalou 1999).

Here we propose that the critical shift from sensory to symbolic representation occurs at the level of the morphological unit. A morpheme, the smallest unit of meaning or structure in language (Aronoff & Fudeman 2022), encodes the meaning of a word, its grammatical function, and its syntactic behavior. For example, the morphological breakdown of the English word *arguments* would comprise *argue-ment-s*: The root morpheme *argue* relates to the core meaning of the word and contains most of the semantic information, the derivational suffix *-ment* is a functional morpheme and serves to indicate the part of speech (noun) of the root, and the inflectional suffix *-s* provides syntactic information that this noun is plural.

Morphological processing has been studied quite extensively in cognitive neuroscience. The operations associated with morphological units have been proposed to be segmentation (identify the morphemes), look-up (extract the appropriate semantic and syntactic features), and composition (combine the features to form a complex representation) (Gwilliams 2020).

One method that researchers have adopted to explore auditory segmentation of morphological structure uses an information theoretic approach (Brodbeck et al. 2018, Gaston & Marantz 2018, Gwilliams & Davis 2022, Gwilliams & Marantz 2015). The approach is to quantify, for

Review in Advance. Changes may still occur before final publication.

Downloaded from www.annualreviews.org. Stanford University Libraries (ar-224679) IP: 171.66.13.59 On: Fri, 25 Oct 2024 15:37:33

each phoneme of the speech input, how much information has been provided to help identify the morphological unit being said. This can be computed under two metrics: surprisal and entropy. Surprisal quantifies how likely a given phoneme is based upon the preceding phonemes within the morphological constituent (Gwilliams & Davis 2022). If a phoneme is less likely, it has higher information content than if it were more likely. Entropy quantifies how certain a given morphological outcome is based on the phonological sequence up to that point. If the morphological outcome is very certain—for example, if the sequence so far is consistent with only one outcome, as in *avalan-*, then the entropy is low (Shannon 1948).

This phoneme-by-phoneme incremental processing is very akin to the Cohort model of speech perception (Marslen-Wilson 1975), which posits an incremental segmentation process for the activation and ultimate identification of lexical items. Lexical and sublexical representations are activated by the phoneme sequence and deactivated if subsequent acoustic or contextual information is inconsistent with them (Marslen-Wilson 1987, Marslen-Wilson & Tyler 1980, Marslen-Wilson & Welsh 1978). Modeling neural responses as a function of each phonemic input is akin to modeling each incremental process hypothesized by the Cohort model.

Having segmented the speech signal for morphological units, the system needs to perform a look-up operation on those units and thereby identify their semantic and syntactic properties. This occurs through the activation of semantic and syntactic features in relative proportion to the likelihood of their occurrence in speech (Gaston & Marantz 2018). Each word comprises three pieces—the root, derivations, and inflections—and depending on the type of morpheme being "looked up," the features extracted are different.

A root morpheme carries semantic properties and represents the core meaning of a given word—for instance, the *poet* in *poetical* or the *appear* in *disappeared*. Accessing the semantic properties of root morphemes has been linked to the STG, the middle temporal gyrus, and the angular gyrus (Binder et al. 2000, Friederici 2012, Hickok & Poeppel 2007, Indefrey & Levelt 2004). Extracting these semantic representations, and its link to conceptual knowledge, is a key component of language processing (Poeppel et al. 2012), and root morpheme access can be equated with accessing the semantic features of a word.

By contrast, a derivational morpheme refers to a constituent (e.g., the *-ic* or *-al* in *poetical*) that determines the part of speech of the word (e.g., noun, verb, adjective), and an inflectional morpheme refers to a constituent that provides additional grammatical information about the word (e.g., *-s* refers to plural, and *-ed* refers to past tense). Access to both types of syntactic information has been linked to the inferior frontal gyrus, which has more broadly been associated with syntactic processing (Carota et al. 2016, Marslen-Wilson & Tyler 2007, Sahin et al. 2009, Whiting et al. 2015).

Finally, having accessed the semantic and syntactic information associated with the morphemes of a word, those representations are combined into a complex whole. This process has primarily been probed by comparing responses to valid and invalid compositional structures (e.g., *farm-er* versus *corn-er*) or by comparing responses that are more or less compatible combinations. Both approaches converge on the orbitofrontal cortex performing the compositional operation by combining the semantic and syntactic properties of the morphemes into a whole word (Fruchter & Marantz 2015, Neophytou et al. 2018, Pylkkänen & McElree 2007).

Breaking down lexical semantic meanings into fundamental bases to construct a quantifiable representational space has long been a challenge. Previous studies tried to define such components using semantic categories or attributes (Tong et al. 2022); however, recent advances in large language models have been particularly advantageous for modeling these complex features of word and multiword meaning in a data-driven manner. By training encoding models with vectorized linguistic representations learned from these models and applying them to neuroimaging and

Review in Advance. Changes may
still occur before final publication.

Downloaded from www.annualreviews.org. Stanford University Libraries (ar-224679) IP: 171.66.13.59 On: Fri, 25 Oct 2024 15:37:33

electrophysiological data (Kell et al. 2018, Li et al. 2023, Yamins & DiCarlo 2016), researchers are able to investigate how lexical semantic meanings are represented in the brain, extending the previous success of this approach in sensory domains (Goldstein et al. 2022, Huth et al. 2016, Kell et al. 2018, Li et al. 2023, Mitchell et al. 2008, Yamins & DiCarlo 2016, Zhang et al. 2020).

For example, the representations in the hidden layers of models such as GloVe (Pennington et al. 2014) and word2vec (Mikolov et al. 2013) encode lexical information, such as the semantic dimensions of the words, and syntactic features, such as parts of speech (Goldstein et al. 2022, Huth et al. 2016, Mitchell et al. 2008, Zhang et al. 2020). From a representational standpoint, these vectors would represent the outcome of the compositional stage of morphological processing.

Insight about the semantic component of morphological and lexical processing has come from semantic impairments, which are characteristic of semantic-variant primary progressive aphasia (svPPA), a progressive aphasia in which the epicenter of neurodegeneration sits primarily in the left-greater-than-right but largely bilateral anterior temporal lobes (ATLs) (Gorno-Tempini et al. 2008, Patterson et al. 2007). Despite being able to accurately repeat words (Leyton et al. 2014) and understand syntactically complex utterances (subject to limited vocabulary; Wilson et al. 2012), individuals with svPPA show immense difficulty understanding word meanings, a symptom that reflects a broader conceptual impairment in which knowledge gradually degrades as a function of taxonomic specificity (i.e., a person with svPPA may understand the word *horse* but not *zebra* and, later, *animal* but not *horse*; Mesulam et al. 2019a, Patterson et al. 2007). While unilateral left anterior temporal lobectomies and strokes do not tend to have a drastic impact on language on their own (Hermann et al. 1991, Tsapkini et al. 2011), the two ATLs working together are theorized to function as a hub in the semantic network, linking domain-specific representations together to build semantic concepts (Patterson & Lambon Ralph 2016, Patterson et al. 2007). Prior work in stroke and primary progressive aphasia (PPA) has also implicated the left angular gyrus as a site relevant for conceptual representation (Geschwind 1965, Price et al. 2015), which may play a role in managing information more specific to thematic roles (i.e., "What do cats do? Cats purr, cuddle, chase…") rather than taxonomic relationships (i.e., "What type of thing are cats? Cats are pets, animals, living things…") (Schwartz et al. 2011).

While comprehension of single lexical items is also often measured in studies of aphasia, lexical comprehension is a relatively coarse metric that belies a complex, multistage process (consisting of auditory, phonemic, syllabic, morphological, and semantic components). Even so, lexical comprehension is actually relatively robust to injury following left hemisphere stroke, particularly as time postinjury increases during the first year of recovery (Selnes et al. 1984, Wilson et al. 2023). Thus, demonstrating a deficit in single-word comprehension does not necessarily provide high granularity with respect to where that deficit occurred. However, when lexical comprehension deficits (not clearly attributable to broader semantic impairment) do occur, they tend to be associated with lesions to the left midposterior STG and sulcus (Hillis et al. 2017, Matchin et al. 2022) or with larger lesions encompassing this territory (Rogalsky et al. 2022, Wilson et al. 2023); reperfusion of this area has also been associated with a restoration of word comprehension abilities (Hillis & Heidler 2002). Cortical stimulation in this region has been shown to impair responsiveness to word comprehension tasks (Lesser et al. 1986) and to induce perceptual deficits at the single-word level (Leonard et al. 2019).

Through Phrasal Composition, the Brain Generates Infinite Meaning from Finite Means

Syntactic phrases, such as noun phrases (e.g., *the fluffy brown and white dog*) and verb phrases (e.g., *cuddled ferociously*), comprise multiple lexical items and together form a systematic structural whole. The ability to flexibly combine and understand others' novel combinations is one of the defining

Review in Advance. Changes may still occur before final publication.

Downloaded from www.annualreviews.org. Stanford University Libraries (ar-224679) IP: 171.66.13.59 On: Fri, 25 Oct 2024 15:37:33

features of human language (Hauser et al. 2002). Yet, compositionality at the phrasal and sentence levels is one of the hardest aspects of language to study, and so it remains an elusive part of language processing.

Understanding how phrases are processed involves considering the processes by which lexical items are combined to create higher-order structure and meaning. Meaning is derived not just from single words but from the combinatorial processes that allow the emergence of complex meaning and relations. In theory, the compositional operations that govern the combination of morphemes to form words are the same as the operations that combine words to form sentences (Matushansky & Marantz 2013, Punske 2023).

To generate sentential meaning, words need to be combined under the syntactic rules that govern their combination. This often means linking words that do not occur in adjacent order; for example, in *the scientist smiled as the magnitude of her discovery sank in*, the pronoun *her* needs to be linked to the prior referent *scientist* to be correctly understood. This adds an important component of difficulty because the system needs to maintain past words and concepts to appropriately combine them with future words and concepts, therefore relying on parallel processing of multiple inputs over time.

Composition, at its core, begins with the presence of two words that need to be combined, such as adjective–noun minimal phrases like *red boat*. Bemis & Pylkkänen (2011) used MEG to contrast neural responses to a noun (e.g., *boat*) that was preceded either by an adjective (e.g., *red*, thus forming a minimal compositional phrase *red boat*) or by a short, meaningless character string (e.g., *rcxn*). Responses to the noun were stronger in the left anterior temporal lobe (LATL) when it was part of a phrase than when it was not (Bemis & Pylkkänen 2013, Westerlund & Pylkkänen 2014).

Is this neural response driven by the syntactic operation that combines minimal phrases, or is it a neural marker of the complexity of conceptual representation being created? In subsequent studies, Pylkkänen and colleagues additionally modified the specificity of the adjective and noun to assess whether this changed the result. Zhang & Pylkkänen (2015) found that using more specific nouns (such as replacing *boat* with *canoe*) decreased the magnitude of response in the LATL, whereas replacing a generic modifier with a more specific one (e.g., replacing *meat* in *meat stew* with *lamb*) strengthened the effect. Because in all cases the same adjective–noun minimal pair is being used, this effect cannot be explained by a syntactic operation that combines adjectives and nouns regardless of their content. Nor can it be explained by the overall specificity (or frequency) of the constituent words. Overall, it suggests that the operations that support basic conceptual structure building are performed in the LATL, which is incremental in its construction of composite concepts.

To investigate how the brain may process compositional meaning, studies have compared reading or listening to sentences with reading or listening to lists of unconnected words. Neuroimaging studies have shown that this comparison elicits a significant difference across many brain regions associated with language processing, including the bilateral anterior temporal poles, bilateral lateral temporal regions, and left hemisphere frontal regions (the inferior frontal gyrus and middle frontal gyrus) (Stowe et al. 1998, Vandenberghe et al. 2002). Using techniques capable of resolving the timing of increased activity, studies have shown that the difference between these two conditions grows over the course of stimulus presentation, possibly indexing construction of meaning over time (Fedorenko et al. 2016, Pallier et al. 2011). This is not to say that each lexical item must be processed serially as it is perceived. Brain activation in language-selective regions is insensitive to some word order swaps that render the sentence ungrammatical but not indecipherable (Mollica et al. 2020), and thus the operations performed on the linguistic information that result in comprehension may be robust to speech errors in word order.

Review in Advance. Changes may still occur before final publication.

Downloaded from www.annualreviews.org. Stanford University Libraries (ar-224679) IP: 171.66.13.59 On: Fri, 25 Oct 2024 15:37:33

While some theories of syntax in the brain suggest that different types of syntactic operations occur in different spatial locations (Caplan et al. 2016), current theories suggest that the key distinction may be between regions predominantly engaged in producing syntactic utterances and those engaged in understanding syntax and its rules (Matchin & Hickok 2020). These theories implicate temporoparietal regions including the left middle temporal gyrus and the neighboring superior temporal sulcus in the role of understanding syntax. Individuals with strokes in these regions struggle with tasks that rely on syntactic comprehension, including sentence–picture matching for semantically reversible sentences (e.g., "The girl washes the boy" versus "The boy washes the girl"; Thothathiri et al. 2012), passive sentences (e.g., "The boy was hugged by the girl"; Caplan et al. 2016), and syntactically noncanonical sentences (e.g., "The carrot that the small rabbit ate is in the garden"; Rogalsky et al. 2018); they additionally struggle to make accurate sentence grammaticality judgments (Wilson & Saygın 2004). Findings in PPA resulting from temporoparietal damage corroborate these results (Amici et al. 2007, Mesulam et al. 2019b). For a recent detailed review regarding syntax in the brain, the interested reader is referred to Matchin & Hickok 2020).

A promising future direction for understanding composition at the level of phrases and sentences is the use of large language models to derive candidate compositional representations (Brown et al. 2020, Devlin et al. 2018). In this case, the representation of *dog* might be different between the contexts *the big angry dog* and *the shiny porcelain dog*, thereby embedding the contextual meaning into the meanings of the individual words. This feature of incorporating a context window has been used to investigate the timescale of information integration (Jain & Huth 2018, Keshishian et al. 2021) and also to model predictive processes during natural speech processing (Goldstein et al. 2022, Huth et al. 2016, Mitchell et al. 2008, Zhang et al. 2020).

In addition, researchers have used multimodal language models, which have a joint optimization function to link captions to the correct images. They have found that compared with models that operate on text only, the visually grounded models learn semantic representations that are more embedded in the physical realization of the object (e.g., animate, inanimate) and better aligned with human intuition and behavior (Zhang et al. 2021). Those semantic grounding models are also offered as computational approaches to explicitly test hypotheses regarding the semantic hub and embodiment in the cortical representations (Tomasello et al. 2018).

Highly Parallel, Highly Redundant

One important theme that emerges from our review is that processing across different hierarchical levels of language unfolds in a highly parallel manner. That is, many processes happen at the same time, in the same spatial region of the brain, and across multiple levels of representation simultaneously. These processing streams are not independent; rather, they engage in constant interaction to mutually generate the most likely interpretation given the dependency structure of language across its levels of representation. We believe that this parallel nature of processing is precisely what makes the system robust against noise, ambiguity, and variability in the sensory speech signal. We also believe that this allows the system to be more efficient because a given region of the brain has access to multiple levels of representation that can serve to resolve noise or ambiguity without having to wait for several rounds of feedforward/feedback loops across regions to reach the correct interpretation.

For example, knowing that a word is likely a noun given the sentence context influences the likelihood of what sounds will occur in that word (a "downward influence" toward the sensory input) as well as the likely meaning of that word (an "upward influence" toward the symbolic referent). Such mutual constraints guide interpretations of the current language input (i.e., "What

Review in Advance. Changes may still occur before final publication.

Downloaded from www.annualreviews.org. Stanford University Libraries (ar-224679) IP: 171.66.13.59 On: Fri, 25 Oct 2024 15:37:33

is happening now?") in addition to influencing interpretation of subsequent inputs (i.e., "What is likely to happen next?") and even influencing the interpretation of previous inputs (i.e., "Do I need to update what I think happened earlier?"). This dynamic system encourages stability and flexibility in processing by ensuring access to multiple formats of representation, from multiple time steps, at the same time.

This observation is in stark contrast to serial bottom-up processing architectures, whereby a given process does not begin until the previous one has finished through to completion (Gaskell & Marslen-Wilson 1997). Such serial processing models typically assume modularity in spatial organization of function, which our review also indicates is misleading—a given region of the brain, such as the STG, encodes multiple speech features, including phonetic contrasts, speaker relative pitch, lexical stress, and statistical likelihood. The serial, modular view is therefore not supported by our overview of the literature. Instead, representations of speech properties, and the operations that generate and manipulate those representations, exist within a highly parallel spatiotemporal regime (McClelland et al. 1987), which can look both forward and backward in time.

A second important theme that emerges from our review is that neural representations are highly redundant. This redundancy stems from two sources. One, perhaps underappreciated, component of language is that its levels of structure are highly correlated and have predictable autocorrelation properties. In practice, that means that the redundant information across levels, thanks to statistical regularity, can be (and is) leveraged by the brain to constrain the ultimate solution. The second is that the brain compounds upon the inherent redundancy of the language input by generating many similar but slightly different representations of the input, and it maintains all of them for an extended period of time. Therefore, "efficient processing" from the brain's perspective does not mean being representationally frugal; it means being representationally greedy. Any and all information that is available to the system is used by the system, in addition to the brain generating highly redundant formats of representations from the input, which emphasize or minimize particular aspects of the signal. We believe that this extensive redundancy is what allows the language system to perform flexibly in different scenarios for different purposes, where each scenario or task may rely upon different formats of representation.

CONCLUSION AND FUTURE STEPS

In this article, we have reviewed a broad body of research to synthesize how the human brain transforms sound into meaning in the service of speech comprehension. By organizing this endeavor into identifying the representations, operations, and information flow, we believe that it is possible to make tangible progress toward this ambitious goal.

One pressing future direction comes from the observation that much less is understood about the processes implemented at the semantic and syntactic levels than at the acoustic and phonological levels, and much less is understood about the operations that the brain applies compared with the representations it generates. We believe that this can be partly attributed to the analytical tools available. Representations can be hypothesized a priori based upon linguistic or auditory theory; they can be quantified and converted into numbers that can then be correlated with neural responses. Operations, by contrast, exist at the transition between representational states. That means that operations can only be inferred indirectly, by comparing the two representations that are hypothesized to be the input and output of that operation. This is made additionally problematic by the reliance on linear analytical methods, which may be unable to capture the true nonlinear transformation. In a similar vein, much more is known about the phonetic levels of processing, partly because the hypothesis space benefits from its linear proximity to the sensory input, which can be easily described in linguistic and auditory neuroscience terms. Higher-order

Review in Advance. Changes may still occur before final publication.

Downloaded from www.annualreviews.org. Stanford University Libraries (ar-224679) IP: 171.66.13.59 On: Fri, 25 Oct 2024 15:37:33

linguistic structures are many more nonlinear transformations away from the sensory input, and the operations that govern them are also more likely to be nonlinear and to incorporate information across longer time constants. A key step forward in addressing both of these shortcomings will be to implement nonlinear analytical methods to uncover the operations implemented in the system, all the way from sound to meaning. This will likely benefit significantly from deep learning tools—models that have been performance optimized for a particular language task to generate hypotheses of nonlinear operations (e.g., Caucheteux et al. 2022) as well as general nonlinear analytical tools that can be used for signal processing (e.g., Keshishian et al. 2020).

Overall, by taking advantage of the impressive technological advances in analytical methods, computational models, and neural recording techniques, it will be possible to uncover the end-to-end computational architecture that governs higher-order language processing. This will entail powerful naturalistic experimental designs to fully engage the comprehension system and assistance from deep learning to model complex symbolic operations to complete our understanding of language comprehension in the human brain.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We thank Dr. Edward Chang for supporting us in writing this article. It was Dr. Chang's interdisciplinary vision that brought us together as trainees in his lab at the University of California, San Francisco (UCSF), and provided us with an incredible environment to explore and express our ideas. This work was funded by the Whitehall Foundation (grant 2024-08-043 awarded to L.G.), the Bill and Ute Bowes Foundation (funding awarded to Chang Lab), the National Institutes of Health (grant R01 DC012379 awarded to Edward Chang, supporting L.G.; grant U01 NS117765 awarded to Edward Chang, supporting Y.Z. and S.H.; Fellowship F32 DC019531 awarded to T.S.; and Fellowship F32 DC020096 awarded to D.L.), and the National Science Foundation (GRFP grant 2034836 awarded to I.B.-G.).

LITERATURE CITED

Ahissar E, Nagarajan S, Ahissar M, Protopapas A, Mahncke H, Merzenich MM. 2001. Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *PNAS* 98(23):13367–72

Amici S, Brambati SM, Wilkins DP, Ogar J, Dronkers NL, et al. 2007. Anatomical correlates of sentence comprehension and verbal working memory in neurodegenerative disease. *J. Neurosci.* 27(23):6282–90

Aronoff M, Fudeman K. 2022. *What Is Morphology?* Hoboken, NJ: Wiley

Barsalou LW. 1999. Perceptual symbol systems. *Behav. Brain Sci.* 22(4):577–609

Belin P, Zatorre RJ. 2000. 'What', 'where' and 'how' in auditory cortex. *Nat. Neurosci.* 3(10):965–66

Bemis DK, Pylkkänen L. 2011. Simple composition: a magnetoencephalography investigation into the comprehension of minimal linguistic phrases. *J. Neurosci.* 31(8):2801–14

Bemis DK, Pylkkänen L. 2013. Basic linguistic composition recruits the left anterior temporal lobe and left angular gyrus during both listening and reading. *Cerebr. Cortex* 23(8):1859–73

Bhaya-Grossman I, Chang EF. 2022. Speech computations of the human superior temporal gyrus. *Annu. Rev. Psychol.* 73:79–102

Binder JR, Frost JA, Hammeke TA, Bellgowan PS, Springer JA, et al. 2000. Human temporal lobe activation by speech and nonspeech sounds. *Cerebr. Cortex* 10(5):512–28

Blevins J. 1995. The syllable in phonological theory. In *Handbook of Phonological Theory*, Vol. 1, ed. J Goldsmith, pp. 206–44. Chichester, UK: Wiley-Blackwell

Review in Advance. Changes may still occur before final publication.

Downloaded from www.annualreviews.org. Stanford University Libraries (ar-224679) IP: 171.66.13.59 On: Fri, 25 Oct 2024 15:37:33

Blumstein SE, Baker E, Goodglass H. 1977. Phonological factors in auditory comprehension in aphasia. *Neuropsychologia* 15(1):19–30

Boatman D. 2004. Cortical bases of speech perception: evidence from functional lesion studies. *Cognition* 92(1–2):47–65

Boatman D, Hall C, Goldstein MH, Lesser R, Gordon B. 1997. Neuroperceptual differences in consonant and vowel discrimination: as revealed by direct cortical electrical interference. *Cortex* 33(1):83–98

Brent MR. 1999. Speech segmentation and word discovery: a computational perspective. *Trends Cogn. Sci.* 3(8):294–301

Brodbeck C, Hong LE, Simon JZ. 2018. Rapid transformation from auditory to linguistic representations of continuous speech. *Curr. Biol.* 28(24):3976–83.e5

Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, et al. 2020. Language models are few-shot learners. arXiv:2005.14165 [cs.CL]

Buchman AS, Garron DC, Trost-Cardamone JE, Wichter MD, Schwartz M. 1986. Word deafness: one hundred years later. *J. Neurol. Neurosurg. Psychiatry* 49(5):489–99

Caplan D, Michaud J, Hufford R, Makris N. 2016. Deficit-lesion correlations in syntactic comprehension in aphasia. *Brain Lang*. 152:14–27

Carota F, Bozic M, Marslen-Wilson W. 2016. Decompositional representation of morphological complexity: multivariate fMRI evidence from Italian. *J. Cogn. Neurosci.* 28(12):1878–96

Casilio M, Kasdan AV, Schneck SM, Entrup JL, Levy DF, et al. 2024. Situating word deafness within aphasia recovery: a case report. *Cortex* 173:96–119

Caucheteux C, Gramfort A, King J-R. 2022. Deep language algorithms predict semantic comprehension from brain activity. *Sci. Rep.* 12(1):16327

Coffey EBJ, Herholz SC, Chepesiuk AMP, Baillet S, Zatorre RJ. 2016. Cortical contributions to the auditory frequency-following response revealed by MEG. *Nat. Commun.* 7:11070

Devlin J, Chang M-W, Lee K, Toutanova K. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 [cs.CL]

Doelling KB, Arnal LH, Ghitza O, Poeppel D. 2014. Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual parsing. *NeuroImage* 85:761–68

Drullman R, Festen JM, Plomp R. 1994a. Effect of reducing slow temporal modulations on speech reception. *J. Acoust. Soc. Am.* 95(5 Part 1):2670–80

Drullman R, Festen JM, Plomp R. 1994b. Effect of temporal envelope smearing on speech reception. *J. Acoust. Soc. Am.* 95(2):1053–64

Dumanch KA, Poling GL. 2019. Introduction to the audiological evaluation: case-based applications to patients with skull base disease. *J. Neurol. Surg. B* 80(2):111–19

Fedorenko E, Scott TL, Brunner P, Coon WG, Pritchett B, et al. 2016. Neural correlate of the construction of sentence meaning. *PNAS* 113(41):E6256–62

Friederici AD. 2012. The cortical language circuit: from auditory perception to sentence comprehension. *Trends Cogn. Sci.* 16(5):262–68

Fruchter J, Marantz A. 2015. Decomposition, lookup, and recombination: MEG evidence for the full decomposition model of complex visual word recognition. *Brain Lang*. 143:81–96

Gaskell MG, Marslen-Wilson WD. 1997. Integrating form and meaning: a distributed model of speech perception. *Lang. Cogn. Process.* 12(5–6):613–56

Gaston P, Marantz A. 2018. The time course of contextual cohort effects in auditory processing of category-ambiguous words: MEG evidence for a single "clash" as noun or verb. *Lang. Cogn. Neurosci.* 33(4):402–23

Geschwind N. 1965. Disconnexion syndromes in animals and man. I. *Brain* 88(2):237–94

Goldstein A, Zada Z, Buchnik E, Schain M, Price A, et al. 2022. Shared computational principles for language processing in humans and deep language models. *Nat. Neurosci.* 25(3):369–80

Gorno-Tempini ML, Brambati SM, Ginex V, Ogar J, Dronkers NF, et al. 2008. The logopenic/phonological variant of primary progressive aphasia. *Neurology* 71(16):1227–34

Gwilliams L. 2020. How the brain composes morphemes into meaning. *Philos. Trans. R. Soc. Lond. B* 375(1791):20190311

Review in Advance. Changes may still occur before final publication.

Downloaded from www.annualreviews.org. Stanford University Libraries (ar-224679) IP: 171.66.13.59 On: Fri, 25 Oct 2024 15:37:33

Gwilliams L, Davis MH. 2022. Extracting language content from speech sounds: the information theoretic approach. In *Speech Perception*, ed. LL Holt, JE Peelle, AB Coffin, AN Popper, RR Fay, pp. 113–39. Cham, Switz.: Springer

Gwilliams L, King J-R, Marantz A, Poeppel D. 2022. Neural dynamics of phoneme sequences reveal position-invariant code for content and order. *Nat. Commun.* 13(1):6606

Gwilliams L, Linzen T, Poeppel D, Marantz A. 2018. In spoken word recognition, the future predicts the past. *J. Neurosci.* 38(35):7585–99

Gwilliams L, Marantz A. 2015. Non-linear processing of a linear speech stream: the influence of morphological structure on the recognition of spoken Arabic words. *Brain Lang.* 147:1–13

Hamilton LS, Edwards E, Chang EF. 2018. A spatial map of onset and sustained responses to speech in the human superior temporal gyrus. *Curr. Biol.* 28(12):1860–71.e4

Hamilton LS, Oganian Y, Hall J, Chang EF. 2021. Parallel and distributed encoding of speech across human auditory cortex. *Cell* 184(18):4626–39.e13

Hauser MD, Chomsky N, Fitch WT. 2002. The faculty of language: What is it, who has it, and how did it evolve? *Science* 298(5598):1569–79

Hermann BP, Seidenberg M, Haltinerr A, Wyler AR. 1991. Mood state in unilateral temporal lobe epilepsy. *Biol. Psychiatry* 30(12):1205–18

Hickok G, Poeppel D. 2007. The cortical organization of speech processing. *Nat. Rev. Neurosci.* 8(5):393–402

Hillis AE, Heidler J. 2002. Mechanisms of early aphasia recovery. *Aphasiology* 16(9):885–95

Hillis AE, Rorden C, Fridriksson J. 2017. Brain regions essential for word comprehension: drawing inferences from patients. *Ann. Neurol.* 81(6):759–68

Holdcroft D. 1991. *Saussure: Signs, System and Arbitrariness*. Cambridge, UK: Cambridge Univ. Press

Huth AG, de Heer WA, Griffiths TL, Theunissen FE, Gallant JL. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* 532(7600):453–58

Indefrey P, Levelt WJM. 2004. The spatial and temporal signatures of word production components. *Cognition* 92(1–2):101–44

Jain S, Huth A. 2018. Incorporating context into language encoding models for fMRI. In *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, ed. S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, R Garnett. Red Hook, NY: Curran. **https://proceedings.neurips.cc/paper_files/paper/2018/file/f471223d1a1614b58a7dc45c9d01df19-Paper.pdf**

Johnson JCS, Jiang J, Bond RL, Benhamou E, Requena-Komuro M-C, et al. 2020. Impaired phonemic discrimination in logopenic variant primary progressive aphasia. *Ann. Clin. Trans. Neurol.* 7(7):1252–57

Johnson K, Sjerps MJ. 2021. Speaker normalization in speech perception. In *The Handbook of Speech Perception*, pp. 145–76. Hoboken, NJ: Wiley

Kell AJE, Yamins DLK, Shook EN, Norman-Haignere SV, McDermott JH. 2018. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron* 98(3):630–44.e16

Keshishian M, Akbari H, Khalighinejad B, Herrero JL, Mehta AD, Mesgarani N. 2020. Estimating and interpreting nonlinear receptive field of sensory neural responses with deep neural network models. *eLife* 9:53445

Keshishian M, Norman-Haignere S, Mesgarani N. 2021. Understanding adaptive, multiscale temporal integration in deep speech recognition systems. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, ed. M Ranzato, A Beygelzimer, Y Dauphinm, PS Liang, J Wortman Vaughan, pp. 24455–67. Red Hook, NY: Curran

Khalighinejad B, Patel P, Herrero JL, Bickel S, Mehta AD, Mesgarani N. 2021. Functional characterization of human Heschl's gyrus in response to natural speech. *NeuroImage* 235:118003

Klatt DH. 1986. Representation of the first formant in speech recognition and in models of the auditory periphery. *Can. Acoust.* 14(3 bis):5–7

Kubanek J, Brunner P, Gunduz A, Poeppel D, Schalk G. 2013. The tracking of speech envelope in the human cortex. *PLOS ONE* 8(1):e53398

Leonard MK, Cai R, Babiak MC, Ren A, Chang EF. 2019. The peri-Sylvian cortical network underlying single word repetition revealed by electrocortical stimulation and direct neural recordings. *Brain Lang.* 193:58–72

Review in Advance. Changes may still occur before final publication.

Downloaded from www.annualreviews.org. Stanford University Libraries (ar-224679) IP: 171.66.13.59 On: Fri, 25 Oct 2024 15:37:33

Leonard MK, Gwilliams L, Sellers KK, Chung JE, Xu D, et al. 2024. Large-scale single-neuron speech sound encoding across the depth of human cortex. *Nature* 626(7999):593–602

Lesser RP, Raudzens P, Lüders H, Nuwer MR, Goldie WD, et al. 1986. Postoperative neurological deficits may occur despite unchanged intraoperative somatosensory evoked potentials. *Ann. Neurol.* 19(1):22–25

Levy DF, Wilson SM. 2020. Categorical encoding of vowels in primary auditory cortex. *Cerebr. Cortex* 30(2):618–27

Ley A, Vroomen J, Hausfeld L, Valente G, De Weerd P, Formisano E. 2012. Learning of new sound categories shapes neural response patterns in human auditory cortex. *J. Neurosci.* 32(38):13273–80

Leyton CE, Savage S, Irish M, Schubert S, Piguet O, et al. 2014. Verbal repetition in primary progressive aphasia and Alzheimer's disease. *J. Alzheimer's Dis.* 41(2):575–85

Li Y, Anumanchipalli GK, Mohamed A, Chen P, Carney LH, et al. 2023. Dissecting neural computations in the human auditory pathway using deep neural networks for speech. *Nat. Neurosci.* 26:2213–25

Liberman AM, Delattre P, Cooper FS. 1952. The role of selected stimulus-variables in the perception of the unvoiced stop consonants. *Am. J. Psychol.* 65(4):497–516

Luo H, Poeppel D. 2007. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron* 54(6):1001–10

Marslen-Wilson WD. 1975. Sentence perception as an interactive parallel process. *Science* 189(4198):226–28

Marslen-Wilson WD. 1987. Functional parallelism in spoken word-recognition. *Cognition* 25(1–2):71–102

Marslen-Wilson WD, Tyler LK. 1980. The temporal structure of spoken language understanding. *Cognition* 8(1):1–71

Marslen-Wilson WD, Tyler LK. 2007. Morphology, language and the brain: the decompositional substrate for language comprehension. *Philos. Trans. R. Soc. Lond. B* 362(1481):823–36

Marslen-Wilson WD, Welsh A. 1978. Processing interactions and lexical access during word recognition in continuous speech. *Cogn. Psychol.* 10(1):29–63

Matchin W, Basilakos A, den Ouden D-B, Stark BC, Hickok G, Fridriksson J. 2022. Functional differentiation in the language network revealed by lesion-symptom mapping. *NeuroImage* 247:118778

Matchin W, Hickok G. 2020. The cortical organization of syntax. *Cerebr. Cortex* 30(3):1481–98

Matushansky O, Marantz AP. 2013. *Distributed Morphology Today: Morphemes for Morris Halle*. Cambridge, MA: MIT Press

McClelland JL, Rumelhart DE, PDP Res. Group. 1987. *Parallel Distributed Processing*, Vol. 2: *Explorations in the Microstructure of Cognition: Psychological and Biological Models*. Cambridge, MA: MIT Press

Mendez MF, Rosenberg S. 1991. Word deafness mistaken for Alzheimer's disease: differential characteristics. *J. Am. Geriatr. Soc.* 39(2):209–11

Mesgarani N, Cheung C, Johnson K, Chang EF. 2014. Phonetic feature encoding in human superior temporal gyrus. *Science* 343(6174):1006–10

Mesulam M-M, Nelson MJ, Hyun J, Rader B, Hurley RS, et al. 2019a. Preferential disruption of auditory word representations in primary progressive aphasia with the neuropathology of FTLD-TDP type A. *Cogn. Behav. Neurol.* 32(1):46–53

Mesulam M-M, Rader BM, Sridhar J, Nelson MJ, Hyun J, et al. 2019b. Word comprehension in temporal cortex and Wernicke area: a PPA perspective. *Neurology* 92(3):e224–33

Mikolov T, Chen K, Corrado G, Dean J. 2013. Efficient estimation of word representations in vector space. arxiv:1301.3781 [cs.CL]

Mitchell TM, Shinkareva SV, Carlson A, Chang K-M, Malave VL, et al. 2008. Predicting human brain activity associated with the meanings of nouns. *Science* 320(5880):1191–95

Moerel M, De Martino F, Formisano E. 2012. Processing of natural sounds in human auditory cortex: tonotopy, spectral tuning, and relation to voice sensitivity. *J. Neurosci.* 32(41):14205–16

Mollica F, Siegelman M, Diachek E, Piantadosi ST, Mineroff Z, et al. 2020. Composition is the core driver of the language-selective network. *Neurobiol. Lang.* 1(1):104–34

Neophytou K, Manouilidou C, Stockall L, Marantz A. 2018. Syntactic and semantic restrictions on morphological recomposition: MEG evidence from Greek. *Brain Lang.* 183:11–20

Oganian Y, Bhaya-Grossman I, Johnson K, Chang EF. 2023. Vowel and formant representation in the human auditory speech cortex. *Neuron* 111(13):2105–18.e4

Review in Advance. Changes may still occur before final publication.

Downloaded from www.annualreviews.org. Stanford University Libraries (ar-224679) IP: 171.66.13.59 On: Fri, 25 Oct 2024 15:37:33

Oganian Y, Chang EF. 2019. A speech envelope landmark for syllable encoding in human superior temporal gyrus. *Sci. Adv.* 5(11):eaay6279

Pallier C, Devauchelle A-D, Dehaene S. 2011. Cortical representation of the constituent structure of sentences. *PNAS* 108(6):2522–27

Patterson K, Lambon Ralph MA. 2016. The hub-and-spoke hypothesis of semantic memory. In *Neurobiology of Language*, ed. G Hickok, SL Small, pp. 765–75. Amsterdam: Academic

Patterson K, Nestor PJ, Rogers TT. 2007. Where do you know what you know? The representation of semantic knowledge in the human brain. *Nat. Rev. Neurosci.* 8(12):976–87

Pennington J, Socher R, Manning C. 2014. Glove: global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), ed. A Moschitti, B Pang, W Daelemans, pp. 1532–43. Stroudsburg, PA: Assoc. Comput. Linguist.

Poeppel D. 2001. Pure word deafness and the bilateral processing of the speech code. *Cogn. Sci.* 25(5):679–93

Poeppel D, Emmorey K, Hickok G, Pylkkänen L. 2012. Towards a new neurobiology of language. *J. Neurosci.* 32(41):14125–31

Price AR, Bonner MF, Peelle JE, Grossman M. 2015. Converging evidence for the neuroanatomic basis of combinatorial semantics in the angular gyrus. *J. Neurosci.* 35(7):3276–84

Punske JP. 2023. *Morphology: A Distributed Morphology Introduction*. Hoboken, NJ: Wiley

Pylkkänen L, McElree B. 2007. An MEG study of silent meaning. *J. Cogn. Neurosci.* 19(11):1905–21

Ray S, Maunsell JHR. 2011. Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. *PLOS Biol.* 9(4):e1000610

Robson H, Grube M, Lambon Ralph MA, Griffiths TD, Sage K. 2013. Fundamental deficits of auditory perception in Wernicke's aphasia. *Cortex* 49(7):1808–22

Rogalsky C, Basilakos A, Rorden C, Pillay S, LaCroix AN, et al. 2022. The neuroanatomy of speech processing: a large-scale lesion study. *J. Cogn. Neurosci.* 34(8):1355–75

Rogalsky C, LaCroix AN, Chen K-H, Anderson SW, Damasio H, et al. 2018. The neurobiology of agrammatic sentence comprehension: a lesion study. *J. Cogn. Neurosci.* 30(2):234–55

Sahin NT, Pinker S, Cash SS, Schomer D, Halgren E. 2009. Sequential processing of lexical, grammatical, and phonological information within Broca's area. *Science* 326(5951):445–49

Schwartz MF, Kimberg DY, Walker GM, Brecher A, Faseyitan OK, et al. 2011. Neuroanatomical dissociation for taxonomic and thematic knowledge in the human brain. *PNAS* 108(20):8520–24

Selnes OA, Niccum N, Knopman DS, Rubens AB. 1984. Recovery of single word comprehension: CT-scan correlates. *Brain Lang.* 21(1):72–84

Shannon CE. 1948. A mathematical theory of communication. *Bell Syst. Tech. J.* 27(3):379–423

Shannon RV, Zeng FG, Kamath V, Wygonski J, Ekelid M. 1995. Speech recognition with primarily temporal cues. *Science* 270(5234):303–4

Simon JZ, Commuri V, Kulasingham JP. 2022. Time-locked auditory cortical responses in the high-gamma band: a window into primary auditory cortex. *Front. Neurosci.* 16:1075369

Stowe LA, Broere CA, Paans AM, Wijers AA, Mulder G, et al. 1998. Localizing components of a complex task: sentence processing and working memory. *NeuroReport* 9(13):2995–99

Thothathiri M, Kimberg DY, Schwartz MF. 2012. The neural basis of reversible sentence comprehension: evidence from voxel-based lesion symptom mapping in aphasia. *J. Cogn. Neurosci.* 24(1):212–22

Tomasello R, Garagnani M, Wennekers T, Pulvermüller F. 2018. A neurobiologically constrained cortex model of semantic grounding with spiking neurons and brain-like connectivity. *Front. Comput. Neurosci.* 12:00088

Tong J, Binder JR, Humphries C, Mazurchuk S, Conant LL, Fernandino L. 2022. A distributed network for multimodal experiential representation of concepts. *J. Neurosci.* 42(37):7121–30

Tsapkini K, Vindiola M, Rapp B. 2011. Patterns of brain reorganization subsequent to left fusiform damage: fMRI evidence from visual processing of words and pseudowords, faces and objects. *Neuroimage* 55(3):1357–72

Vandenberghe R, Nobre AC, Price CJ. 2002. The response of left temporal cortex to sentences. *J. Cogn. Neurosci.* 14(4):550–60

Westerlund M, Pylkkänen L. 2014. The role of the left anterior temporal lobe in semantic composition vs. semantic memory. *Neuropsychologia* 57:59–70

Review in Advance. Changes may still occur before final publication.

Whiteford KL, Kreft HA, Oxenham AJ. 2020. The role of cochlear place coding in the perception of frequency modulation. *eLife* 9:58468

Whiting C, Shtyrov Y, Marslen-Wilson W. 2015. Real-time functional architecture of visual word recognition. *J. Cogn. Neurosci.* 27(2):246–65

Wilson SM, Entrup JL, Schneck SM, Onuscheck CF, Levy DF, et al. 2023. Recovery from aphasia in the first year after stroke. *Brain* 146(3):1021–39

Wilson SM, Galantucci S, Tartaglia MC, Gorno-Tempini ML. 2012. The neural basis of syntactic deficits in primary progressive aphasia. *Brain Lang*. 122(3):190–98

Wilson SM, Saygın AP. 2004. Grammaticality judgment in aphasia: Deficits are not specific to syntactic structures, aphasic syndromes, or lesion sites. *J. Cogn. Neurosci.* 16(2):238–52

Yamins DLK, DiCarlo JJ. 2016. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* 19(3):356–65

Yi HG, Leonard MK, Chang EF. 2019. The encoding of speech sounds in the superior temporal gyrus. *Neuron* 102(6):1096–110

Zatorre RJ, Evans AC, Meyer E, Gjedde A. 1992. Lateralization of phonetic and pitch discrimination in speech processing. *Science* 256(5058):846–49

Zhang L, Pylkkänen L. 2015. The interplay of composition and concept specificity in the left anterior temporal lobe: an MEG study. *NeuroImage* 111:228–40

Zhang Y, Han K, Worth R, Liu Z. 2020. Connecting concepts in the brain by mapping cortical representations of semantic relations. *Nat. Commun.* 11(1):1877

Zhang Y, Choi M, Han K, Liu Z. 2021. Explainable semantic space by grounding language to vision with cross-modal contrastive learning. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, ed. M Ranzato, A Beygelzimer, Y Dauphin, PS Liang, J Wortman Vaughan, pp. 18513–26. Red Hook, NY: Curran

Downloaded from www.annualreviews.org. Stanford University Libraries (ar-224679) IP: 171.66.13.59 On: Fri, 25 Oct 2024 15:37:33

Review in Advance. Changes may
still occur before final publication.