NeuroImage 171 (2018) 62–74

Contents lists available at ScienceDirect

NeuroImage

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

Convergence of spoken and written language processing in the superior temporal sulcus

Stephen M. Wilson [a,*], Alexa Bautista [b], Angelica McCarron [b]

[a] Department of Hearing and Speech Sciences, Vanderbilt University Medical Center, Nashville, TN, USA
[b] Department of Speech, Language, and Hearing Sciences, University of Arizona, Tucson, AZ, USA

ARTICLE INFO

Keywords:
Dorsal bank
Functional MRI
Language comprehension
Narrative
Ventral bank

ABSTRACT

Spoken and written language processing streams converge in the superior temporal sulcus (STS), but the functional and anatomical nature of this convergence is not clear. We used functional MRI to quantify neural responses to spoken and written language, along with unintelligible stimuli in each modality, and employed several strategies to segregate activations on the dorsal and ventral banks of the STS. We found that intelligible and unintelligible inputs in both modalities activated the dorsal bank of the STS. The posterior dorsal bank was able to discriminate between modalities based on distributed patterns of activity, pointing to a role in encoding of phonological and orthographic word forms. The anterior dorsal bank was agnostic to input modality, suggesting that this region represents abstract lexical nodes. In the ventral bank of the STS, responses to unintelligible inputs in both modalities were attenuated, while intelligible inputs continued to drive activation, indicative of higher level semantic and syntactic processing. Our results suggest that the processing of spoken and written language converges on the posterior dorsal bank of the STS, which is the first of a heterogeneous set of language regions within the STS, with distinct functions spanning a broad range of linguistic processes.

Introduction

Spoken and written language take very different perceptual forms. The speech waveform enters the auditory system as a continuous stream containing spectro-temporal cues to phonemes that the listener must segment and map onto phonological word forms. In contrast, written language enters the brain in the form of patterns of light on the retina; the reader makes saccades to fixate on successive chunks of text, identifies letters, and maps them onto orthographic word forms. In either case, the final goal is the same: to derive a conceptual representation of meaning. But to get to that endpoint, there are also processing stages that are largely independent of the input modality, for instance, accessing the meanings of words from their forms, combining their meanings according to the syntactic structure of the utterance, and so on. These basic observations suggest a "Y-shaped" model of spoken and written language processing, in which two distinct modality-specific streams of processing converge at some point onto a modality-neutral common processing stream, which ultimately yields an abstract representation of meaning.

The cortical pathways involved in the early, modality-specific stages of processing of both spoken and written language are quite well understood. For spoken language, primary and higher level auditory areas in Heschl's gyrus and on the dorsal and lateral surfaces of the superior temporal gyrus (STG) carry out spectro-temporal analysis of the auditory signal (Binder et al., 1996; Formisano et al., 2003; Mesgarani et al., 2014; see Moerel et al., 2014 for review). For written language, a hierarchy of occipital and ventral temporal regions in the ventral visual stream code increasingly complex and abstract visual features of the letter string (Binder and Mohr, 1992; Cohen et al., 2000; Vinckier et al., 2007; Dehaene and Cohen, 2011). The cortical correlates of the conceptual representations that constitute the endpoint of language comprehension are also increasingly well understood. This semantic system comprises a network of brain regions including the middle temporal gyrus (MTG), anterior temporal lobe, angular gyrus, and inferior frontal gyrus (IFG) (Geschwind, 1965; Patterson et al., 2007; Binder et al., 2009; Visser et al., 2012; Huth et al., 2016).

What is less clear is the functional neuroanatomy of the intervening processes and representations, including precisely how and where the processing of spoken and written language converges. Several functional imaging studies have shown that neural activity common to the processing of spoken and written language is localized to the superior temporal sulcus (STS), predominantly in the left hemisphere (Spitsyna et al., 2006; Jobard et al., 2007; Lindenberg and Scheef, 2007; Berl et al.,

* Corresponding author. Department of Hearing and Speech Sciences, Vanderbilt University Medical Center, 1215 21st Ave S, MCE 8310, Nashville, TN 37232, USA.
E-mail address: stephen.m.wilson@vanderbilt.edu (S.M. Wilson).

https://doi.org/10.1016/j.neuroimage.2017.12.068
Received 2 August 2017; Received in revised form 15 December 2017; Accepted 20 December 2017
Available online 23 December 2017

2010). Moreover, the STS is similarly modulated by rate and intelligibility in both modalities (Vagharchakian et al., 2012), and the time courses of STS responses to the same linguistic material in spoken and written form are remarkably similar (Regev et al., 2013). Taken together, these studies suggest that spoken and written language processing converge in the STS.

While this finding is a vital first step, it leaves many important questions unanswered, because the STS is a not a unitary structure (Liebenthal et al., 2014). Rather, it is a deep sulcus containing a great expanse of neural tissue. Studies in non-human primates have shown that the STS contains numerous subdivisions with distinct cytoarchitectonic properties and connectivity profiles (Jones and Powell, 1970; Seltzer and Pandya, 1978). In the domain of language, the STS has been implicated in a heterogeneous range of processes, covering the gamut of stages from sublexical processing of speech (Liebenthal et al., 2005; Möttönen et al., 2006; Uppenkamp et al., 2006; Turkeltaub and Coslett, 2010; Liebenthal et al., 2014), to representation of phonological word forms (Okada and Hickok, 2006), to semantic and syntactic processing (Scott et al., 2000; Davis and Johnsrude, 2003; Friederici et al., 2009; Wilson et al., 2016). In both the spoken and written modalities, regions in the STS are sensitive to manipulation of lower level (van Atteveldt et al., 2004) and higher level (Xu et al., 2005; Jobard et al., 2007) aspects of language processing.

To better understand how spoken and written language processing streams converge in the STS, it is first necessary to clarify the functional parcellation of the STS with respect to language. This undertaking faces two main challenges: one linguistic, and the other anatomical. The first challenge is that language processing generally involves seamless and integrated computations at multiple levels: phonological or orthographic, lexical, semantic, syntactic and so on. In functional imaging studies, even the most ingenious contrasts between conditions (e.g. Scott et al., 2000) often end up entailing multiple differences between conditions, at more than one level of representation (Binder, 2000). In the present study, we addressed this challenge by investigating not only contrasts between carefully matched intelligible and unintelligible spoken and written inputs, but also by quantifying neural responses to the unintelligible inputs themselves (Woodhead et al., 2011). Models of spoken and written language processing (e.g. McClelland and Rumelhart, 1981; McClelland and Elman, 1986; Taylor et al., 2013) make clear predictions about the extent to which different kinds of unintelligible inputs should drive different levels of linguistic processing. Furthermore, we used searchlight multi-voxel pattern analysis (MVPA; Kriegeskorte et al., 2006) to identify brain regions that can distinguish between different inputs by means of distributed patterns of signal change, even if they show the same overall level of activation (Haxby et al., 2001; Kamitani and Tong, 2005).

The second challenge to parcellating the STS is anatomical: the dorsal and ventral banks of the STS are, by nature, in close physical proximity to one another, and functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) generally lack the spatial resolution to distinguish between activity on the two banks of the sulcus. While fMRI has higher spatial resolution than PET, the blood oxygen level-dependent (BOLD) signal that is the basis of most fMRI studies is more sensitive to signal changes in draining veins than in gray matter itself (Bandettini and Wong, 1997; Lai et al., 1993; Menon et al., 1993), and medium-sized draining veins run through the STS, as they do through all major sulci. Therefore, in typical fMRI studies, activations in the STS are localized to the veins that run through the sulcus, which are downstream of the location(s) where neural activity is occurring, and are therefore somewhat uninformative with regard to the specific site of the neural activity (Wilson, 2014). To address this challenge, we employed several strategies to maximize spatial resolution. First, small voxels were acquired, and no spatial smoothing was applied. Second, a breath-holding task in a separate run was used to estimate and correct for voxelwise differences in cerebrovascular reactivity (CVR) (Bandettini and Wong, 1997; Cohen et al., 2004; Handwerker et al., 2007; Thomason et al., 2007; Murphy et al., 2011; Wilson, 2014); this effectively de-emphasizes signal from veins, which have very high CVR (Wilson, 2014). Third, veins were identified on susceptibility-weighted imaging (SWI), and masked out. Fourth, intersubject normalization was carried out with the large-deformation DARTEL registration algorithm (Ashburner, 2007), which aligns specific structures across participants with exceptional accuracy (Klein et al., 2009). Taken together, these methodological choices were intended to facilitate the identification of distinct patterns of responses to intelligible and unintelligible spoken and written inputs on the dorsal and ventral banks of the STS, in order to further our understanding of how spoken and written language processing streams converge in the STS.

Materials and methods

Participants

Sixteen healthy participants of a wide range of ages took part in the study (mean age = 57 years; range = 21–81 years; 9 females; 1 left-hander and 2 ambidextrous). No participant reported any history of neurological disorders. All participants gave written informed consent, and the study was approved by the institutional review board at the University of Arizona.

Narrative comprehension paradigm

Each participant completed two ($N = 5$) or three ($N = 11$) narrative comprehension runs. There were five conditions: listening to spoken narrative segments ("Spoken"), listening to backwards spoken narrative segments ("Backwards"), reading written narrative segments ("Written"), quasi-reading scrambled written narrative segments ("Scrambled"), and no stimulus ("Rest"). Each run comprised 15 segments per condition, presented in pseudorandom order. A sparse sampling protocol was used, with a repetition time (TR) of 9500 ms and an acquisition time (TA) of 2269 ms, leaving 7231 ms silence between successive acquisitions. Two initial volumes were acquired and discarded, and then one image was acquired after each stimulus or rest period, for a total of 75 volumes per run.

The narrative was the beginning of an audiobook recording of the novel *Hope Was Here* by Joan Bauer, read by Jenna Lamua (Bauer, 2004). The narrative was split into segments at pauses such that each segment was as long as possible up to 7 s (occasionally, slightly longer segments were extracted, then reduced to 7 s by shortening internal pauses). The mean length of the segments was 5656 ms ± 1012 (SD) ms.

In the Spoken narrative condition (Fig. 1A), each narrative segment was presented centered in the silent interval between scans, such that the peak of a typical hemodynamic response to the segment would coincide with acquisition of the subsequent image.

The Backwards narrative condition (Fig. 1B) was the same, except that the segments were played in reverse, rendering them unintelligible. Note that backwards speech contains partial phonemic information. In particular, monophthongal vowels are not greatly affected by reversal, and many consonants also retain their identities. Previous research has shown that naive transcription of backwards words is considerably better than chance (Binder et al., 2000), supporting the notion that backwards speech carries phonemic information; it seems plausible that phonemic information could also be extracted from backwards sentences. Models of spoken word comprehension generally posit that representations of phonemes are mapped onto lexical modes by a spreading activation mechanism (McClelland and Elman, 1986). From this perspective, because it contains recognizable phonemes, the Backwards condition would be expected to activate brain regions involved in phonemic representation of spoken inputs. Moreover, due to spreading activation between phonemic and lexical representations, the Backwards condition should also activate brain regions involved in representation of lexical nodes, even though no lexical nodes will ultimately be selected. Because no lexical nodes are selected, brain regions involved in semantic