LANGUAGE, COGNITION AND NEUROSCIENCE
2020, VOL. 35, NO. 5, 573–582
https://doi.org/10.1080/23273798.2018.1499946

Routledge
Taylor & Francis Group

REVIEW ARTICLE

OPEN ACCESS Check for updates

The revolution will not be controlled: natural stimuli in speech neuroscience

Liberty S. Hamilton[a,b] and Alexander G. Huth[c,d]

[a]Communication Sciences & Disorders, Moody College of Communication, The University of Texas at Austin, Austin, USA; [b]Department of Neurology, Dell Medical School, The University of Texas at Austin, Austin, USA; [c]Department of Neuroscience, The University of Texas at Austin, Austin, USA; [d]Department of Computer Science, The University of Texas at Austin, Austin, USA

ABSTRACT

Humans have a unique ability to produce and consume rich, complex, and varied language in order to communicate ideas to one another. Still, outside of natural reading, the most common methods for studying how our brains process speech or understand language use only isolated words or simple sentences. Recent studies have upset this *status quo* by employing complex natural stimuli and measuring how the brain responds to language as it is used. In this article we argue that natural stimuli offer many advantages over simplified, controlled stimuli for studying how language is processed by the brain. Furthermore, the downsides of using natural language stimuli can be mitigated using modern statistical and computational techniques.

ARTICLE HISTORY
Received 21 February 2018
Accepted 3 July 2018

KEYWORDS
Natural language; encoding models; fMRI; ECoG; EEG

A fundamental goal in neuroscience is to discover how the human brain understands and produces language. The methods used to study processes in the human brain have advanced considerably over the past decades. Advancements in neuroimaging and neural recording technologies have made it possible to measure brain responses with higher fidelity and spatio-temporal resolution, and modern analysis techniques have made it possible to analyze larger and more complex datasets. Yet many—if not most—experimental designs in neurolinguistics are still rooted in the techniques of the past: comparing brain responses to isolated words or simplified sentences. One alternative is to perform experiments using natural language stimuli, with connected sentences that approximate or draw directly from language as it is used in everyday life. Outside of neuroscience, highly natural approaches have already found use in conversation analysis (CA), where natural social conversations are analysed qualitatively (Kendrick, 2017; Schegloff, Koshik, Jacoby, & Olsher, 2002). Natural stimuli have also been used widely in studies concerned with the neural processes that underlie reading behaviours (Kliegl, Dambacher, Dimigen, Jacobs, & Sommer, 2012). But for studies that probe how the brain understands language or processes speech, natural stimuli have found only limited use. A few recent studies have shown that conclusions based on simplified or highly controlled language stimuli may not apply to data collected using natural language, or that similar conclusions can be reached more efficiently using natural language (Huth, de Heer, Griffiths, Theunissen, & Gallant, 2016; Lerner, Honey, Silbert, & Hasson, 2011; Wehbe et al., 2014).

Recent efforts to use natural language stimuli in neuroscience closely echo debates that occurred in visual neuroscience over the past 20 years. That field was dominated for decades by an experimental approach in which tightly controlled visual stimuli were used to study receptive field properties of neurons in visual cortex. This was successful in characterising many properties of the visual cortex, including retinotopic representations of the visual field, ocular dominance columns, the receptive field properties of simple and complex cells in the visual pathway, and more. Yet over time it became clear that many effects assumed to be universal were actually highly dependent on the tightly controlled stimuli, and were diminished or absent in experiments that used natural visual stimuli (David, Vinje, & Gallant, 2004). Recently many visual neuroscience experiments have begun to use more natural stimuli either to construct or test models of visual processing (Geisler, Perry, Super, & Gallogly, 2001; Kay, Naselaris, Prenger, & Gallant, 2008; Nishimoto & Gallant, 2011; Rao & Ballard, 1999).

The changes in vision neuroscience were spurred largely by technology. Both measurements of brain

CONTACT Alexander G. Huth huth@cs.utexas.edu, alex.huth@gmail.com

activity and computational resources have improved dramatically in recent years, making it feasible to fit complex computational models to brain data. These same technologies are available to language researchers, so there has never been a better time to start the natural language revolution. In this paper we will describe some procedures for analyzing data from experiments that use natural language stimuli, then we will provide statistical arguments for widespread adoption of these procedures. We will then present some caveats of the experimental methods and suggestions for moving forward, and will close by discussing a few key results that have come from natural language experiments.

What constitutes natural language in perception?

Before discussing how natural language stimuli are used in neuroscience research it is important to clarify what we mean by the term. In "real life" situations (Matusz, Dikker, Huth, and Perrodin, 2018), using natural language would involve listening, speech production, turn taking, and many other communicative signals, including non-linguistic utterances and gestures. However, here we focus only on the more limited domain of natural language perception. We propose that naturalness of a stimulus lies along a spectrum and can be gauged by answering three questions. First, is this a stimulus that a person might reasonably be exposed to outside of an experimental setting? Second, does the stimulus appear in the same context as it would in real life? Third, is the subject's motivation for perceiving and understanding the stimulus particular to the experimental setting, or is it a motivation that the subject would feel in real life? Here, we consider several types of language stimuli through the lens of these questions.

Isolated words

Many experiments use isolated word stimuli, and ask subjects to perform specific tasks such as rating whether it is a word or non-word (Binder, Desai, Graves, & Conant, 2009). Clearly humans are able to understand and reason about isolated words, but this type of stimulus fails at least two questions of naturalness. While single-word utterances are not uncommon in real-world settings, they are most often generated by pre-grammatical children, adults who are answering questions ("Yes" or "No"), or as imperative statements (e.g. "Duck!") (Greenfield, 1978). These natural single-word utterances have pragmatic contexts that are missing in an experimental setting. Furthermore, experimental tasks such as judging words vs. nonwords are far from ethological.

Isolated sentences

Many experiments use isolated sentences as stimuli (Anderson et al., 2016; Hamilton, Edwards, & Chang, 2018; Just, Wang, & Cherkassky, 2017; Mesgarani, Cheung, Johnson, & Chang, 2014). These stimuli are often drawn from real-world sources, so they clearly pass the first criterion. However, it is relatively uncommon in real life to hear sentences such as "The couple laughed at dinner" divorced from any context whatsoever. Because subjects have little intrinsic motivation to comprehend or process decontextualised sentences, these experiments often include tasks, such as deciding whether a sentence is grammatical. Such tasks are a form of motivation that is not common in real life. Isolated sentences thus are clearly more natural than isolated words, but they still not completely natural.

Complete narrative stories

Several recent experiments have used complete narrative stories or book chapters as stimuli (de Heer, Huth, Griffiths, Gallant, & Theunissen, 2017; Honey, Thompson, Lerner, & Hasson, 2012; Huth et al., 2016; Lerner et al., 2011; Wehbe et al., 2014). These stimuli are drawn from real-world sources, so they automatically pass the first question. Complete narratives also ensure that every sentence occurs in a natural context, passing the second question. And drawing stimuli from popular entertainment sources such as *Harry Potter* (Wehbe et al., 2014) or *The Moth* (de Heer et al., 2017; Huth et al., 2016; Lerner et al., 2011) helps address the issue of motivation, since many people choose to consume those stimuli voluntarily for no reason other than to comprehend them. This type is the most natural language stimulus that has been used in laboratory neuroscience experiments. (However, recent work has begun to examine language and social interaction in a natural setting, going a step further toward truly natural neuroscience (Bevilacqua et al., 2018).)

Before closing this section we note that how one interprets the questions we posed about naturalness depends on the goal of the research. For example, in a study designed to probe phonological representations, narrative stories might be considered no more natural than isolated sentences. It is unlikely that phoneme representations rely on more context than could be provided by a sentence, and the question of motivation becomes less well-defined for phoneme processing than it is for language understanding. But even though isolated sentences may be sufficiently natural to study phoneme representations without any penalty, there may be little downside to using more contextualised stimuli such as narrative stories, as long as they contain

sufficient variation in phonetic content. Similarly, a study designed to probe high-level semantic and syntactic representations might present narrative stories using a controlled presentation paradigm such as serial visual presentation (SVP) instead of natural reading. This could still be considered natural language for the purpose of the study, since the high-level processes in question are likely insensitive to the presentation method.

Statistical methods for natural stimulus experiments

One profound issue that has undoubtedly delayed the widespread adoption of natural language in neuroscience experiments is that the data often cannot be analysed using the same techniques that are used for traditional controlled experiments. Traditional experiments are typically constructed such that the data can be analysed using statistical techniques such as *t*-tests, *F*-tests, or ANOVAs. However, these methods are poorly suited for most natural stimulus experiments because they cannot control for confounding or correlated variables (Although there have been instances where these techniques have been used to great effect with carefully constructed narratives (Saxe & Kanwisher, 2003).) Thus, it is common for natural language experiments to employ other statistical techniques. Here we will briefly describe three techniques that have been used to study how the brain processes natural language stimuli: encoding models, unsupervised dimensionality reduction, and inter-subject correlation.

Encoding models are quantitative, mathematical models that are designed to predict brain responses based on the stimuli that elicited them (Naselaris, Kay, Nishimoto, & Gallant, 2011). These models typically have a number of free parameters that are estimated using one dataset, termed the "training dataset". The free parameters are then fixed, and this encoding model is used to predict brain responses in a "validation dataset" that was not used during parameter estimation. Encoding model performance can then be assessed by comparing predicted and actual responses in the held-out validation dataset (using, for example, Pearson correlation). Variants of this technique have been used for decades for auditory electrophysiology (Aertsen & Johannesma, 1981; Ahrens, Linden, & Sahani, 2008; Carruthers, Natan, & Geffen, 2013; Theunissen, Sen, & Doupe, 2000). More recently, encoding models have been applied to a number of different language-related questions in fMRI (de Heer et al., 2017; Huth et al., 2016; Kandylaki et al., 2016; Mitchell et al., 2008; Wehbe et al., 2014), ECoG (Berezutskaya, Freudenburg, Güçlü, van Gerven, & Ramsey, 2017; Cheung, Hamilton, Johnson, & Chang, 2016; Hamilton et al., 2018; Holdgraf et al., 2017; Hullett, Hamilton, Mesgarani, Schreiner, & Chang, 2016; Mesgarani et al., 2014; Tang, Hamilton, & Chang, 2017), and EEG (Crosse, Di Liberto, Bednar, & Lalor, 2016; Di Liberto, O'Sullivan, & Lalor, 2015). In fMRI, this technique is sometimes known as voxel-wise modelling (VM) (Huth et al., 2016); it may also be called a multivariate temporal response function (Crosse, Di Liberto, Bednar, et al., 2016) or spectrotemporal receptive field models (STRFs) in the auditory domain (Aertsen & Johannesma, 1981).

There are many different forms of encoding models (Holdgraf et al., 2017; Wu, David, & Gallant, 2006), but the most common variant is the "linearised model". In linearised models, features are extracted from the stimuli using any available technique (e.g. hand-labeling, automatic labelling, or unsupervised statistical methods). These features are then combined by a linear regression model that attempts to predict brain responses. Linearised models have demonstrated the existence of strong spectrotemporal and phonetic feature representations in superior temporal gyrus (de Heer et al., 2017; Di Liberto et al., 2015; Hullett et al., 2016; Mesgarani et al., 2014) and motor cortex (Cheung et al., 2016), independent representations of pitch intonation information, speaker identity, and sentence identity in bilateral superior temporal gyrus (Tang et al., 2017), and semantic representations across a wide swath of cortex (Huth et al., 2016). They have also shown how spectrotemporal, articulatory, and semantic information contribute to the generation of neural signals in different cortical areas (de Heer et al., 2017), and how feature representations may be modulated by attention (Fritz, Elhilali, David, & Shamma, 2007; Mesgarani & Chang, 2012; O'Sullivan, Reilly, & Lalor, 2015), intelligibility (Holdgraf et al., 2016; Khoshkhoo, Leonard, Mesgarani, & Chang, 2018), or behavioural context (David, 2017).

While encoding models provide a good way of testing the relative contributions of one set of features versus another in predicting brain responses, the features of interest are not always known *a priori*. Thus, researchers may want to use unsupervised methods to analyze their neural data that do not impose a predefined set of features. For example, unsupervised dimensionality reduction techniques such as convex non-negative matrix factorisation (cNMF, (Ding, Li, & Jordan, 2010)) can be used find spatiotemporal patterns of brain activity that can later be correlated with specific stimulus features (Hamilton et al., 2018). In Hamilton et al., for example, the researchers played naturally spoken sentences to patients with implanted intracranial electrodes covering language-related cortical areas including

superior temporal gyrus, and then used cNMF to find patterns of neural activity that were consistent across participants listening to the same natural sentences. An advantage of this unsupervised analysis is that they could discover features of interest without imposing a specific hypothesis about the main drivers of neural activity. They found that neural activity was grouped into two major response types: "onset" related activity, found more posteriorly, and "sustained" activity, found anteriorly. When correlating these with known acoustic-phonetic features, they found the "onset" electrodes responded strongly at sentence and phrase onsets, and that phonetic feature responses in this area were highly adapting and context-sensitive. The "sustained" electrodes, in contrast, were active throughout the sentence and did not show distinct responses to phonetic content at the beginning vs. at the middle of the sentence. Perhaps most importantly, this "onset" and "sustained" distinction is not easily appreciated using simpler, shorter stimuli such as consonant–vowel (CV) syllables. With CV syllables (and also with short, single words), it is difficult to appreciate the larger response at the onset, since there is no continuous information following the syllable, and effectively every stimulus is an onset. Again, this research points to the importance of using natural stimuli in order to uncover these response types that would only be seen in longer, more complex stimuli.

A third technique that has been applied to natural stimulus experiments is inter-subject correlation (ISC) (Hasson, Nir, Levy, Fuhrmann, & Malach, 2004; Honey et al., 2012; Lerner et al., 2011; Silbert, Honey, Simony, Poeppel, & Hasson, 2014). In this technique brain responses are recorded from multiple subjects while they are presented with natural stimuli. For each brain area, response time courses from different subjects are correlated to determine whether the responses are similar across subjects. This method provides an estimate of the "signal-to-noise ratio" for a particular stimulus in each brain area, while being agnostic to the temporal profile of the response. This is somewhat different from the information that is offered by more traditional contrast-based approaches, in that ISC is a function of the temporal pattern of the response in addition to response size. Sensitivity to temporal patterns means that ISC can be used to study responses at timescales up to minutes, whereas traditional approaches are generally limited to relatively short stimulus blocks. In (Lerner et al., 2011), for example, ISC was used to compare the reliability of stimulus-evoked responses across stimuli that had varying amounts of temporal context from single words to a complete 10-minute narrative. This showed that the stimuli with the most temporal context—the most natural stimuli—evoked responses across a much larger expanse of cortex than the other stimuli.

Like unsupervised methods, ISC does not require the specification of a feature space, making this technique simple to apply. It also has the advantage of being more computationally simple than unsupervised methods such as cNMF. However, ISC has the same downside as unsupervised methods: it can tell you that a brain area responds consistently to a stimulus, but not why. ISC also has a further constraint in that it assumes that brain areas are matched accurately between subjects, usually using anatomical co-registration. If this co-registration is faulty, then one would observe low ISC even if the underlying brain responses were highly similar. However, techniques such as hyperalignment (Haxby et al., 2011) promise to correct this issue.

Statistical arguments for natural stimuli

Generalisability

A scientific result is only useful if it is *generalisable*, in the sense that similar effects should be observed using other stimuli or other subjects. As mentioned previously, results based on simplified, non-natural stimuli such as words or isolated sentences may fail to generalise. If our goal is to understand how the brain processes language in ethological settings, then we should be concerned with whether experimental results generalise to all natural language stimuli. Here experiments based on natural language stimuli have a clear advantage. The only barrier to generalisation is that results obtained from one domain of natural language may not apply to other domains. This drawback can be mitigated by sampling stimuli broadly.

Failures to generalise from simple stimuli to complex natural stimuli arise from the fact that brain responses at any moment are a nonlinear function of current and previous stimuli (Leonard, Baud, Sjerps, & Chang, 2016; Lewicki & Arthur, 1996; Williamson, Ahrens, Linden, & Sahani, 2016). Thus it is difficult, for example, to predict the response to a series of words knowing only the response to each word in isolation. The only way to understand the nonlinear functions implemented by the brain, and thus to predict responses to natural language, is to learn from responses to natural language stimuli that include those nonlinear contextual effects.

Effect size

One point that is frequently raised in statistical critiques of neuroscience and psychology is that too much attention is paid to statistical significance and too little to

effect size (Button et al., 2013; Carver, 1993). Effect size provides information about the importance of an effect, while significance provides information about its reliability. However, it is difficult to compare effect sizes across experiments that use different methodologies, different types of stimuli, different numbers of subjects, or involve different areas of the brain. For example, the superior temporal gyrus (STG) has been shown to represent many different features present in speech, but comparing these representations across studies is nearly impossible. This problem is made especially acute by the small stimulus sets used in many experiments, which likely inflate significance and could alter effect size (Westfall, Nichols, & Yarkoni, 2016).

Natural stimuli provide a solution to this problem by providing a gold-standard measure—the fraction of variance in responses to natural stimuli that can be explained, or *natural effect size*—that captures both the validity and importance of an effect. Every hypothesised effect can be framed as a model of how the brain responds to natural language in perception or production. For example, one might hypothesise that abstract and concrete words elicit different brain responses. In an experiment employing natural stimuli, each word in a narrative could be tagged as "abstract" or "concrete", yielding two features that could be used to estimate predictive encoding models. One would then use the fitted model to predict responses on a new natural language dataset, and compute the natural effect size as the fraction of variance in the new dataset predicted by the model. This value could then be directly compared to other effects—such as phoneme content, syntactic properties, or prosodic contours—because each effect can be instantiated as a model of the same, natural dataset.

Experimental efficiency

In most controlled experiments, the hypothesis that is being tested is built into the experimental design. This is typically done by constructing experimental conditions that differ on one variable, such as whether they contain a semantically incongruent word, but are matched on as many other variables as possible. This type of experiment provides high statistical power for addressing the hypothesis of interest, but is essentially useless for testing most other hypotheses.

Natural language experiments are typically designed without a specific hypothesis in mind, but instead sample stimuli broadly from some domain. Natural stimuli differ on many variables, but the amount of variation is limited, and some variables will have more variation in natural stimuli than others. For example, most natural language stimuli contain relatively few errors of semantic congruity. This type of experiment thus has reduced power for testing any particular hypotheses relative to a controlled experiment designed for that hypothesis. However, because it is not designed with a hypothesis in mind, a single natural language experiment can be used to test many different hypotheses. This renders experiments employing natural stimuli more efficient. Furthermore, using the same dataset to examine different hypotheses makes it possible to compare and disentangle the contributions of different variables, such as semantic, phonological, and spectrotemporal features (de Heer et al., 2017), or pitch vs. speaker vs. sentence identity (Hamilton et al., 2018; Hullett et al., 2016; Mesgarani et al., 2014; Tang et al., 2017).

Statistical caveats in natural language experiments

While natural language experiments offer many advantages over traditional experimental paradigms, they are not without pitfalls. Here we discuss a number of these issues and ways to circumvent or minimise them.

Stimulus correlations

In a controlled experimental setting, one attempts to eliminate any confounding stimulus features that may be correlated with the hypothesised effect. For example, in an experiment designed to study how the brain responds to semantically incongruent words, subjects may be presented with pairs of sentences that are identical except for a single word, which is either congruent or incongruent with the rest of the sentence. Because all but one of the words in each sentence are fixed across the two conditions, this design reduces the correlation between the variable of interest (whether the word is semantically congruent) and other properties of the stimuli. In a natural language experiment this type of control is, by definition, impossible. Any hypothesised effect in natural language will invariably be correlated with one or more confounding variables. This limits the effectiveness of simple statistical tools such as t-tests for analyzing data from natural stimulus experiments.

However, encoding models embody the solution to the problem of confounding variables: regression analysis. Entering confounding variables into a regression analysis along with the variables of interest can disentangle the contributions of each to the total recorded brain response. This is possible for any confounding variable, as long as that variable can be quantified and included in the model. Furthermore, the degree to which any two variables are confounded, and thus the uncertainty

in how to apportion variance between them, can be quantified by examining posterior distributions in a Bayesian linear regression setting.

In most situations the virtual control offered by regression can suffice to de-confound the variables of interest. However it is possible that a variable of interest and confounding variable are too highly correlated for regression to be effective. One strategy to deal with this situation is to simply increase the size of the stimulus set. However, this can be expensive and time-consuming. Another possible strategy is to redesign the natural stimulus set by oversampling natural stimuli that break the undesired correlation. However, this has the potential to distort the parameters learned in regression models.

Low power for rare variables

One caveat for natural stimulus experiments is that they have low power for inferring how the brain responds to variables that naturally have a low rate of occurrence. For example, the phoneme "ʒ" (ARPABET = "ZH", as in "vi**s**ion", "collu**s**ion", and "mira**g**e") appears much less frequently in natural English speech than most other phonemes. Having fewer occurrences leads to noisier estimates of how the brain responds, since there are fewer responses that can be averaged. One strategy for dealing with this problem is to oversample natural stimuli that contain the rare variables. This strategy was taken in the TIMIT acoustic-phonetic database, where "phonetically diverse" sentences were designed to sample the distribution of all phonemes in English, including rare phonemes and phoneme combinations, more often than they would normally be seen in truly natural speech (Garofolo, Lamel, Fisher, Fiscus, & Pallett, 1993). However, similar to oversampling uncorrelated stimuli, this has the potential to distort regression parameters. The safer, albeit more expensive, option is to simply increase the size of the stimulus.

Limited stimulus domain

Any natural stimulus experiment will invariably draw stimuli from within a particular domain, such as autobiographical stories (Huth et al., 2016) or a fictional narrative (Di Liberto et al., 2015; Wehbe et al., 2014). Depending on the particular properties of that domain, this has the potential to limit generalizability. Besides the obvious solution of using stimuli from as many domains as possible, there is little that can be done to mitigate this problem. Thus it is important that this issue is acknowledged and discussed. In addition, for some experimental situations, such as working with children or in patient populations where time is limited or data collection is otherwise more difficult, natural stimulus selection must be done carefully in order to maximise the probability of being able to fit the models of interest.

Splitting the difference: manipulating natural language stimuli

While using completely naturalistic stimuli has some disadvantages as described above, researchers have also made significant progress by taking naturalistic stimuli and manipulating them in specific ways in order to address specific questions. These questions have included the relative separability of acoustic-phonetic and prosodic information (Tang et al., 2017), how comprehension affects language representation (Adank & Devlin, 2010; Broderick, Anderson, Di Liberto, Crosse, & Lalor, 2018; Peelle, Gross, & Davis, 2013), how degrading stimuli by adding noise influences specific feature representations (Di Liberto, Crosse, & Lalor, 2018; Ding & Simon, 2013), how natural stimulus statistics influence the ability to segregate simultaneous speech streams (Popham, Boebinger, Ellis, Kawahara, & McDermott, 2018), how temporal structure affects speech processing (Lerner et al., 2011; Overath, McDermott, Zarate, & Poeppel, 2015), and how prior knowledge affects representations of previously incomprehensible stimuli (Davis & Johnsrude, 2007; Di Liberto, Lalor, & Millman, 2018; Holdgraf et al., 2016; Khoshkhoo et al., 2018), among many others. In each of these studies, natural language stimuli were systematically manipulated in order to address a specific question. In Tang et al., for example, natural sentences were manipulated so that only the pitch information was changed in order to disentangle the contributions of phonetic feature information that were invariant to pitch. This manipulation effectively decorrelates some aspects of the stimuli that might be more correlated in a purely natural experiment. In Popham et al., natural sentences were manipulated to change a single harmonic such that utterances were perceived as "inharmonic", which resulted in increased difficulty in separating speech streams with this type of information. In Lerner et al. and Overath et al., the amount of temporal information presented to the auditory system was manipulated either by scrambling natural stimuli at different time scales, or by constructing "sound quilts" from natural stimuli with varying segment length. In Khoshkoo et al. and Holdgraf et al., the researchers showed that a synthesised incomprehensible stimulus can be repeated after presentation of the original natural speech, and top-down expectations of the speech percept will both "fill in" what the listener hears as well as influence brain activity that reflects this. These studies complement both pure natural

language and highly controlled experiments and provide the best of both worlds when a specific question is to be addressed.

What have we learned from natural language studies?

Although the majority of research on how the brain processes language continues to use controlled stimuli, a number of studies over the past few decades have used natural language. While these studies explored many different questions, here we wish to highlight three overarching themes that natural language studies have helped elucidate.

First, one major difference between results from natural language and traditional studies is the anatomical extent of language-related activity in the brain. Statistical contrasts employed in controlled studies typically identify a few brain areas at a time, with some language-specific areas such as Wernicke's and Broca's showing up in many studies. Studies may even restrict their analyses to predefined regions of interest (ROIs), usually in order to maximise statistical power. Natural language studies, on the other hand, often reveal much more widespread responses to language (Huth et al., 2016; Lerner et al., 2011). One important study used inter-subject correlation to determine which areas respond strongly to single words, sentences, or coherent narratives (Lerner et al., 2011). That study showed that narratives elicited the most widespread responses, followed by sentences, and then words. Another study showed that many or most of the brain areas that respond to narratives are actually selective for certain categories of words, or semantic domains (Huth et al., 2016). Similar results can only be obtained from controlled studies by way of meta-analyses that combine information from many studies (Binder et al., 1997, 2009). And while the brain areas identified as responding to language in these studies include many that may be considered "multi-purpose" rather than "language-specific" (Fedorenko, Behr, & Kanwisher, 2011), these results still highlight an important difference between data collected using controlled and natural stimuli.

A second and related point is that studies using natural language seem to elicit responses that are considerably less left-lateralized than those seen in traditional studies. The idea that the left cerebral hemisphere is specialised for language and the right is not has been pervasive in language neuroscience since the nineteenth century (Berwick, Friederici, Chomsky, & Bolhuis, 2013; Geschwind, 1970; Wada & Rasmussen, 1960; Wernicke, 1970). Empirical studies have generally shown a left hemisphere dominance for language perception and production (see review in (Price, 2010)). On the other hand, research employing simple word stimuli originally localised prosodic and "emotional" content of speech to the right hemisphere (Schirmer & Kotz, 2006; Van Lancker & Fromkin, 1973). Recent experiments have instead shown that natural language comprehension (and coordination between speech perception and production) involves bilateral networks (Belin, Zatorre, Lafaille, Ahad, & Pike, 2000; Cogan et al., 2014; Hamilton et al., 2018; Huth et al., 2016; Jung-Beeman, 2005; Lerner et al., 2011; Obleser, Eisner, & Kotz, 2008), and that prosodic information is also processed bilaterally (Tang et al., 2017). Notably, one of the largest differences between experiments using natural stimuli vs. simple word or syllable stimuli is that the extent of activation and the involvement of higher order cortical areas is much greater when stimuli are meaningful and have long term structure.

Third, natural language experiments that employed voxelwise encoding models have proven highly efficient at answering—in one experiment—questions that would have required many traditional experiments. In (Wehbe et al., 2014), data collected while subjects read chapters from *Harry Potter* was used to explore representations of syntax, named entities, and semantic meaning. This study was able to compare the relative contributions of these different variables to brain responses because all the variables were present in the same set of natural stimuli. In (de Heer et al., 2017), data collected while subjects listened to stories from *The Moth Radio Hour* were used to compare spectrotemporal, phonemic, and semantic representations. In (Huth et al., 2016), the same data as in (de Heer et al., 2017) were used to compare representations of dozens of different semantic categories. In (Crosse, Di Liberto, & Lalor, 2016), EEG was collected while participants viewed natural audiovisual stimuli with varying levels of noise to examine how visual speech enhances auditory representations. In (Broderick et al., 2018), the same data as in (Crosse, Di Liberto, & Lalor, 2016) were used to look at the effect of intelligibility of these stimuli on semantic representations in the brain. The same could be done using traditional experimental designs and analyses, but likely have required many times more data, and would require that the categories were pre-specified. These examples show how efficient natural language experiments can be for exploring many different questions using a single dataset.

Conclusion

In this paper we have outlined a number of issues surrounding the use of natural language stimuli in neuroscience experiments. Natural language has a number of

clear advantages over traditional controlled stimuli, and a number of downsides. Until recently, it was impossible to take advantage of natural language stimuli and difficult to mitigate their shortcomings. However, explosive growth in computational resources, dataset size, and data quality have brought these techniques within reach of all researchers. In particular, sophisticated computational and statistical techniques that were all but impossible 20 years ago can be done on a laptop today. We want to emphasise that the goal of this paper is not to cast aspersions on research that was done in the past, but to encourage language researchers to think hard about whether a natural language approach could benefit their research questions.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by Burroughs Wellcome Fund.

References

Adank, P., & Devlin, J. T. (2010). On-Line plasticity in spoken sentence comprehension: Adapting to time-compressed speech. *NeuroImage*, *49*(1), 1124–1132.

Aertsen, A. M., & Johannesma, P. I. (1981). The spectro-temporal receptive field. A functional characteristic of auditory neurons. *Biological Cybernetics*, *42*(2), 133–143.

Ahrens, M. B., Linden, J. F., & Sahani, M. (2008). Nonlinearities and contextual influences in auditory cortical responses modeled with multilinear spectrotemporal methods. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *28*(8), 1929–1942.

Anderson, A. J., Binder, J. R., Fernandino, L., Humphries, C. J., Conant, L. L., Aguilar, M., … Raizada, R. D. S. (2017). Predicting neural activity patterns associated with sentences using a neurobiologically motivated model of semantic representation. *Cerebral Cortex*, *27*(9), 4379–4395.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, *403* (6767), 309–312.

Berezutskaya, J., Freudenburg, Z. V., Güçlü, U., van Gerven, M. A. J., & Ramsey, N. F. (2017). Neural tuning to low-level features of speech throughout the Perisylvian Cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *37*(33), 7906–7920.

Berwick, R. C., Friederici, A. D., Chomsky, N., & Bolhuis, J. J. (2013). Evolution, brain, and the nature of language. *Trends in Cognitive Sciences*, *17*(2), 89–98.

Bevilacqua, D., Davidesco, I., Wan, L., Oostrik, M., Chaloner, K., Rowland, J., & Dikker, S. (2018). Brain-to-brain synchrony and learning outcomes vary by student-teacher dynamics: Evidence from a real-world classroom electroencephalography study. *Journal of Cognitive Neuroscience*, 1–11. https://doi.org/10.1162/jocn_a_01274

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.

Binder, J. R., Frost, J. A., Hammeke, T. A., Cox, R. W., Rao, S. M., & Prieto, T. (1997). Human brain language areas identified by functional magnetic resonance imaging. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *17*(1), 353–362.

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, *28*(5), 803–809.e3.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376.

Carruthers, I. M., Natan, R. G., & Geffen, M. N. (2013). Encoding of ultrasonic vocalizations in the auditory cortex. *Journal of Neurophysiology*, *109*(7), 1912–1927.

Carver, R. P. (1993). The case against statistical significance testing, revisited. *The Journal of Experimental Education*, *61* (4), 287–292.

Cheung, C., Hamilton, L. S., Johnson, K., & Chang, E. F. (2016, March). The auditory representation of speech sounds in human motor cortex. *eLife*, *5*, ncbi.nlm.nih.gov. doi:10.7554/eLife.12577

Cogan, G. B., Thesen, T., Carlson, C., Doyle, W., Devinsky, O., & Pesaran, B. (2014). Sensory-motor transformations for speech occur bilaterally. *Nature*, *507*(7490), 94–98.

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The multivariate temporal response function (mTRF) toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience*, *10*, 604.

Crosse, M. J., Di Liberto, G. M., & Lalor, E. C. (2016). Eye can hear clearly now: Inverse effectiveness in natural audiovisual speech processing relies on long-term crossmodal temporal integration. *Journal of Neuroscience*, *36*(38), 9888–9895.

David, S. V. (2017). Incorporating behavioral and sensory context into spectro-temporal models of auditory encoding. *Hearing Research*, *360*, 107–123. doi:10.1016/j.heares.2017.12.021

David, S. V., Vinje, W. E., & Gallant, J. L. (2004). Natural stimulus statistics alter the receptive field structure of v1 neurons. *Journal of Neuroscience*, *24*(31), 6991–7006.

Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hearing Research*, *229*(1–2), 132–147.

de Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L., & Theunissen, F. E. (2017). The hierarchical cortical organization of human speech processing. *The Journal of Neuroscience*, *37* (27), 6539–6557.

Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Cortical measures of phoneme-level speech encoding correlate with the perceived clarity of natural speech. *eNeuro*, *5*(2). doi:10.1523/ENEURO.0084-18.2018

Di Liberto, G. M., Lalor, E. C., & Millman, R. E. (2018, February). Causal cortical dynamics of a predictive enhancement of speech intelligibility. *NeuroImage*, *166*, 247–258.

Di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, *25*(19), 2457–2465.

Ding, C., Li, T., & Jordan, M. I. (2010). Convex and semi-nonnegative matrix factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(1), 45–55.

Ding, N., & Simon, J. Z. (2013). Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *Journal of Neuroscience*, *33*(13), 5728–5735.

Fedorenko, E., Behr, M. K., & Kanwisher, N. (2011). Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences*, *108* (39), 16428–16433.

Fritz, J. B., Elhilali, M., David, S. V., & Shamma, S. A. (2007). Auditory attention—focusing the searchlight on sound. *Current Opinion in Neurobiology*, *17*(4), 437–455.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., & Pallett, D. S. (1993). DARPA TIMIT Acoustic-Phonetic Continous Speech Corpus CD-ROM. NIST Speech Disc 1-1.1. *NASA STI/Recon Technical Report N* 93. adsabs.harvard.edu. Retrieved from http://adsabs.harvard.edu/abs/1993STIN … 9327403G

Geisler, W. S., Perry, J. S., Super, B. J., & Gallogly, D. P. (2001). Edge Co-occurrence in natural images predicts contour grouping performance. *Vision Research*, *41*(6), 711–724.

Geschwind, N. (1970). The organization of language and the brain: Language disorders after brain damage help in elucidating the neural basis of verbal behavior. *Science*, *170* (3961), 940–944.

Greenfield, P. M. (1978). Informativeness, presupposition, and semantic choice in single-word utterances. In N. Waterson & C Snow (Eds.), *Development of communication: Social and pragmatic factors in language acquisition* (pp. 159–166). New York and London: Wiley.

Hamilton, L. S., Edwards, E., & Chang, E. F. (2018). A spatial map of onset and sustained responses to speech in the human superior temporal gyrus. *Current Biology*, *28*(12), 1860–1871.e4. doi:10.1016/j.cub.2018.04.033

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science*, *303*(5664), 1634–1640.

Haxby, J. V., Swaroop Guntupalli, J., Connolly, A. C., Halchenko, Y. O., Conroy, B. R., Ida Gobbini, M., … Ramadge, P. J. (2011). A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, *72*(2), 404–416.

Holdgraf, C. R., de Heer, W., Pasley, B., Rieger, J., Crone, N., Lin, J. J., … Theunissen, F. E. (2016, December). Rapid tuning shifts in human auditory cortex enhance speech intelligibility. *Nature Communications*, *7*. nature.com: 13654

Holdgraf, C. R., Rieger, J. W., Micheli, C., Martin, S., Knight, R. T., & Theunissen, F. E. (2017). Encoding and decoding models in cognitive electrophysiology. *Frontiers in Systems Neuroscience*, *11*(61), 1. doi:10.3389/fnsys.2017.00061

Honey, C. J., Thompson, C. R., Lerner, Y., & Hasson, U. (2012). Not lost in translation: Neural responses shared across languages. *Journal of Neuroscience*, *32*(44), 15277–15283.

Hullett, P. W., Hamilton, L. S., Mesgarani, N., Schreiner, C. E., & Chang, E. F. (2016). Human superior temporal gyrus organization of spectrotemporal modulation tuning derived from speech stimuli. *Journal of Neuroscience*, *36*(6), 2014–2026.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458.

Jung-Beeman, M. (2005). Bilateral brain processes for comprehending natural language. *Trends in Cognitive Sciences*, *9* (11), 512–518.

Just, M. A., Wang, J., & Cherkassky, V. L. (2017, August). Neural representations of the concepts in simple sentences: Concept activation prediction and context effects. *NeuroImage*, *157*, 511–520.

Kandylaki, K. D., Nagels, A., Tune, S., Kircher, T., Wiese, R., Schlesewsky, M., & Bornkessel-Schlesewsky, I. (2016). Predicting "when" in discourse engages the human dorsal auditory stream: An fMRI study using naturalistic stories. *Journal of Neuroscience*, *36*(48), 12180–12191.

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*(7185), 352–355.

Kendrick, K. H. (2017). Using conversation analysis in the Lab. *Research on Language and Social Interaction*, *50*(1), 1–11.

Khoshkhoo, S., Leonard, M. K., Mesgarani, N., & Chang, E. F. (2018, January). Neural correlates of sine-wave speech intelligibility in human frontal and temporal cortex. *Brain and Language*. doi:10.1016/j.bandl.2018.01.007

Kliegl, R., Dambacher, M., Dimigen, O., Jacobs, A. M., & Sommer, W. (2012). Eye movements and brain electric potentials during reading. *Psychological Research*, *76*(2), 145–158.

Leonard, M. K., Baud, M. O., Sjerps, M. J., & Chang, E. F. (2016, December). Perceptual restoration of masked speech in human cortex. *Nature Communications*, *7*. nature.com: 13619

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience*, *31* (8), 2906–2915.

Lewicki, M. S., & Arthur, B. J. (1996). Hierarchical organization of auditory temporal context sensitivity. *The Journal of Neuroscience*, *16*(21), 6987–6998.

Matusz, P. J., Dikker, S., Huth, A. G., & Perrodin, C. (2018). Are we ready for real-world neuroscience? *Journal of Cognitive Neuroscience*, 1. https://doi.org/10.1162/jocn_e_01276

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, *485*(7397), 233–236.

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, *343*(6174), 1006–1010.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, *320*(5880), 1191–1195.

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, *56*(2), 400–410.

Nishimoto, S., & Gallant, J. L. (2011). A three-dimensional spatiotemporal receptive field model explains responses of area MT neurons to naturalistic movies. *Journal of Neuroscience*, *31*(41), 14551–14564.

Obleser, J., Eisner, F., & Kotz, S. A. (2008). Bilateral speech comprehension reflects differential sensitivity to spectral and temporal features. *Journal of Neuroscience*, *28*(32), 8116–8123.

O'Sullivan, J. A., Reilly, R. B., & Lalor, E. C. (2015). *Improved decoding of attentional selection in a cocktail party environment with EEG via automatic selection of relevant independent components*. Conference proceedings: … annual international conference of the IEEE engineering in medicine and biology society. IEEE engineering in medicine and biology

society. Conference 2015 (August). ieeexplore.ieee.org (pp. 5740–5743).

Overath, T., McDermott, J. H., Zarate, J. M., & Poeppel, D. (2015). The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nature Neuroscience*, *18*(6), 903–911.

Peelle, J. E., Gross, J., & Davis, M. H. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral Cortex*, *23*(6), 1378–1387.

Popham, S., Boebinger, D., Ellis, D. P. W., Kawahara, H., & McDermott, J. H. (2018). Inharmonic speech reveals the role of harmonicity in the cocktail party problem. *Nature Communications*, *9*(1), 327.

Price, C. J. (2010, March). The anatomy of language: A review of 100 fMRI studies published in 2009. *Annals of the New York Academy of Sciences*, *1191*, 62–88.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking peopleThe role of the temporo-parietal junction in "theory of mind". *NeuroImage*, *19*(4), 1835–1842.

Schegloff, E., Koshik, I., Jacoby, S., & Olsher, D. (2002). 1. Conversation analysis and applied linguistics. *Annual Review of Applied Linguistics*, *22*, 3–31.

Schirmer, A., & Kotz, S. A. (2006). Beyond the right hemisphere: Brain mechanisms mediating vocal emotional processing. *Trends in Cognitive Sciences*, *10*(1), 24–30.

Silbert, L. J., Honey, C. J., Simony, E., Poeppel, D., & Hasson, U. (2014). Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. *Proceedings of the National Academy of Sciences*, *111*(43), E4687–E4696. doi:10.1073/pnas.1323812111

Tang, C., Hamilton, L. S., & Chang, E. F. (2017). Intonational speech prosody encoding in the human auditory cortex. *Science*, *357*(6353), 797–801.

Theunissen, F. E., Sen, K., & Doupe, A. J. (2000). Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *The Journal of Neuroscience*, *20*(6), 2315–2331.

Van Lancker, D., & Fromkin, V. A. (1973). Hemispheric specialization for pitch and 'tone': Evidence from Thai. *Journal of Phonetics*, *1*(2), 101–109.

Wada, J., & Rasmussen, T. (1960). Intracarotid injection of sodium amytal for the lateralization of cerebral speech dominance. *Journal of Neurosurgery*, *17*(2), 266–282.

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., & Mitchell, T. (2014). Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS One*, *9*(11), e112575–e112575.

Wernicke, C. (1970). The aphasic symptom-complex. *Archives of Neurology*, *22*(3), 280–282.

Westfall, J., Nichols, T. E., & Yarkoni, T. (2016). Fixing the stimulus-as-fixed-effect fallacy in task fMRI. *Wellcome Open Research*, *1*, 23.

Williamson, R. S., Ahrens, M. B., Linden, J. F., & Sahani, M. (2016). Input-Specific gain modulation by local sensory context shapes cortical and thalamic responses to complex sounds. *Neuron*, *91*(2), 467–481.

Wu, M. C.-K., David, S. V., & Gallant, J. L. (2006). Complete functional characterization of sensory neurons by system identification. *Annual Review of Neuroscience*, *29*(1), 477–505.