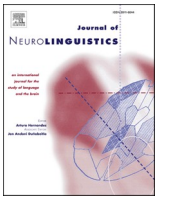
Journal of Neurolinguistics 69 (2024) 101177

Contents lists available at ScienceDirect

Journal of Neurolinguistics

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



Physiological responses and cognitive behaviours: Measures of heart rate variability index language knowledge

Dagmar Divjak [a,b,*], Hui Sun [a,1], Petar Milin [a]

[a] Department of Modern Languages, University of Birmingham, Edgbaston, Ashley Building, Birmingham, B15 2TT, United Kingdom
[b] Department of English Language & Linguistics, University of Birmingham, Edgbaston, Ashley Building, Birmingham, B15 2TT, United Kingdom

A B S T R A C T

Over the past decades, focus has been on developing methods that allow tapping into aspects of cognition that are not directly observable. This includes linguistic knowledge and skills which develop largely without awareness and may therefore be difficult or impossible to articulate. Building on the relation between language cognition and the nervous system, we examine whether Heart Rate Variability (HRV), a cardiovascular measure that indexes Autonomic Nervous System activity, can be used to assess implicit language knowledge. We test the potential of HRV to detect whether individuals possess grammatical knowledge and explore how sensitive the cardiovascular response is.

41 healthy, British English-speaking adults listened to 40 English speech samples, half of which contained grammatical errors. Thought Technology's 5-channel ProComp 5 encoder tracked heart rate via a BVP-Flex/Pro sensor attached to the middle finger of the non-dominant hand, at a rate of 2048 samples per second. A Generalised Additive Mixed Effects Model confirmed a cardiovascular response to grammatical violations: there is a statistically significant reduction in HRV as indexed by NN50 in response to stimuli that contain errors. The cardiovascular response reflects the extent of the linguistic violations, and NN50 decreases linearly with an increase in the number of errors, up to a certain level, after which HRV remains constant.

This observation brings into focus a new dimension of the intricate relationship between physiology and cognition. Being able to use a highly portable and non-intrusive technique with language stimuli also creates exciting possibilities for assessing the language knowledge of individuals from a range of populations in their natural environment and in authentic communicative situations.

1. Introduction

Over the past decades, much effort has been put into developing methods that allow tapping into cognition, and in particular into those aspects of cognition that are not directly observable. These so-called implicit measures infer mental contents from responses on performance-based tasks, thereby making it possible to capture a hypothesized function or process without (conscious self-)assessment (De Houwer, Teige-Mocigemba, Spruyt, & Moors, 2009; Gawronski & De Houwer, 2014). In this respect, implicit measures differ markedly from explicit measures which have been criticized due to their susceptibility to various forms of bias (e.g., response styles, socially desirable responding, impression management; see Rust & Golombok, 2014; Gawronski & Hahn, 2018).

In research on language, the question of which measures to use is of particular importance. Accurately assessing an individual's linguistic abilities, regardless of age and physical or cognitive abilities, is important for many questions pertaining to core areas of life relating to cognition, including brain health. Because the linguistic knowledge of first language (L1) and the associated skills are

* Corresponding author. Department of Modern Languages, University of Birmingham, Edgbaston, Ashley Building, Birmingham, B15 2TT, United Kingdom.

E-mail addresses: d.divjak@bham.ac.uk (D. Divjak), sunh21@cardiff.ac.uk (H. Sun), p.milin@bham.ac.uk (P. Milin).

[1] Present address: School of English, Communication and Philosophy, Cardiff University, John Percival Building, Column Road, Cardiff CF10 3EU, United Kingdom.

https://doi.org/10.1016/j.jneuroling.2023.101177
Received 16 February 2023; Received in revised form 7 September 2023; Accepted 27 September 2023
Available online 12 October 2023

largely implicit, i.e., they have been developed without awareness and are typically deployed without awareness (see Ellis, 2015 for a discussion), this knowledge may therefore be difficult or impossible to articulate explicitly and hence to access using explicit measures. Yet, the main methods theoretical linguists use to assess linguistic knowledge rely on an explicit intermediary: metalinguistic ability. Language users are asked, for example, to judge the grammaticality of an utterance. However, explicitly judging the grammaticality of an utterance is not something language users usually do. Typically, language users produce and process language for communicative purposes, in real time. Judgments, however, expect users to focus on language that has been produced by someone else, and judge it along criteria that are related to the forms used, with a little delay. This taps into language user's metalinguistic ability, i.e., their ability to focus attention on language as an object in and of itself, to reflect upon language, and to evaluate it (Roehr-Brackin, 2018). For this reason, metalinguistic tasks are said to tap into specific resources, in particular cognitive control. This means that, even in healthy populations, individual differences in the ability to isolate the formal dimension of language and retain focus on the formal aspects may affect judgments. Furthermore, these explicit measures suffer additionally from known problems of bias, in the sense that knowledge gained through education regarding which forms are and are not (considered) acceptable may cloud judgments since off-line judgments allow sufficient time for participants who have such knowledge to access and involve it.

To address these issues and enable the reliable detection of linguistic knowledge without reliance on explicit, metalinguistic judgments, research has long relied on reaction times. Reaction times measure the amount of time it takes an individual to respond to a stimulus. The elapsed time is taken as an indicator of the cognitive efficiency in processing the stimulus. Through careful manipulation of properties of the stimulus, it is possible to establish what facilitates or impedes processing. In recent years, and with technological advances, the arsenal of implicit measurements has been expanded with a number of ways to capture physiological responses that correlate with cognitive behaviours but occur automatically and therefore involuntarily (cf. Soares et al. (2015) report that neurophysiological versions of language tests detect cognitive decline more reliably). These physiological responses are obtained using techniques ranging from eye-tracking over electro-encephalography to brain imaging. An added benefit of these techniques is that they track stimulus processing in real time while allowing a look inside the black box of our minds.

The goal of this paper is to zoom in on the relation between language cognition and the autonomic nervous system (ANS), which has so far received less attention, even though data from pupillometry suggests that ANS-related measures offer a promising avenue for studying language cognition. In Section 1.1, we will first survey some of the physiological ways in which the effects of the Autonomic Nervous System manifest. In Section 1.2 we will survey key studies that demonstrate a link between one particular physiological measure, Heart Rate Variability, and cognitive function. In Sections 2 and 3 we will present a study, designed to test whether cardiovascular measures allow us to record an individual's sensitivity to grammatical violations accurately enough to help us assess their language knowledge.

1.1. *The autonomic nervous system and its physiological expression*

Physiological responses are governed by the autonomic nervous system (ANS). The ANS comprises two parts: the sympathetic (SNS) and the parasympathetic (PNS) nervous system. Simply put, the sympathetic nervous system activates the "fight or flight" response during a threat or perceived danger, while the parasympathetic nervous system controls the "rest and digest" or "feed and breed" functions of the body. The sympathetic nervous system, which controls the body's responses to perceived threats, is responsible for increasing heartbeat, tensing muscles, dilating the pupil (to let in more light) and inhibiting saliva secretion, among other things. Due to the short(er) neural pathway of the SNS, it has a faster response time than the PNS which regulates the body's functions while at rest. Together they regulate the involuntary and reflexive functions of the body.

Physiological expressions of the ANS have been linked to aspects of cognition. A widely known use of ANS-related measures in this regard is found in lie detectors: physiological measures (autonomic, electrocortical, or neurovascular) have long been used to detect concealed information in suspects. Although as of yet no physiological profile of lying has been established, ANS measures have been used to successfully differentiate between concealed knowledge (e.g., crime-related knowledge of a guilty person) and absent knowledge (of an innocent person) with remarkable accuracy (Ambach & Gamer, 2018).

In the study of language the desire to replace explicit measures with implicit measures, and record involuntary physiological responses to a stimulus, has led to the reliance on information gleaned from eye movements (Rayner, 1998, 2009). The measurement of pupil dilation or changes in pupil diameter is of particular relevance here as the pupillary response is thought to reflect the combined contributions of the ANS and as such captures some amalgamation of attention, engagement, arousal, anxiety, and effort (Zekveld, Heslenfeld, Johnsrude, Versfeld, & Kramer, 2014; Winn, Wendt, Koelewijn, & Kuchinsky, 2018, pp. 2, 4). Task-evoked pupil responses have been demonstrated in numerous studies (see Beatty, 1982 for an overview): changes in pupil dilation distinguish cognitive tasks that are more or less effortful across a wide variety of domains and pupil dilation amplitude has become a useful measure of task-evoked *resource allocation* (Beatty & Lucero-Wagoner, 2000). At the same time, an increase in pupil dilation therefore also signals arousal, often interpreted as the extent to which an individual *engages* with the task since pupils also dilate with activation of the sympathetic nervous system and inhibition of the parasympathetic system (Loewenfeld, 1993; Steinhauer, Siegle, Condray, & Pless, 2004). However, the neurology[2] behind cognition-mediated pupil responses requires further investigation to be fully understood (see Zekveld et al., 2014 for a first study). Likewise, whether the pupil dilation that occurs on increased load is due to task-demand or

[2] The pupillary response appears to reflect activation of the locus coeruleus (LC) and norepinephrine (NE) system (LC-NE), which has been associated with several cognitive functions such as memory, attention, reward anticipation and decision making (Zekveld et al., 2014). The LC-NE plays an important role in controlling automatic functions and pupil size.

task-engagement (arousal) remains to be determined. Despite this, the technique has seen a considerable uptake in the study of language over the past decade (see Schmidtke, 2018 for an overview) and has been applied across a range of visual (Just & Carpenter, 1993) and auditory (Ben-Nun, 1986) tasks, including in younger (Chapman & Hallowell, 2015) and older (Scherger, 2022) clinical populations.

While the pupil diameter can increase by as much as 3–4 mm, or roughly 120%, when changing from light to dark environments, cognitive task-evoked pupil dilations are much smaller by comparison, in the order of 0.1–0.5 mm, depending on testing conditions and task (Winn et al., 2018). Language-related tasks involving eye-tracking typically expect literate participants with normal or corrected-to-normal vision to view stimuli, be it images or text, on a computer screen. To obtain reliable pupil size measurements for objects the size of letters, a costly, high-precision eye-tracker must be used and participants must have their head fixated in a chin rest. This set-up restricts the researcher's ability to observe individuals from a wide range of populations, regardless of age or disability (affecting vision) and educational background (affecting literacy levels) for longer periods of time in their natural environment and in authentic communicative situations.

In this study, we explore other physiological indicators that can be used to assess language knowledge. We are particularly interested in measures that are governed by the same neuro-physiological system as the pupils and can (now) be reliably recorded using non-invasive and low-cost techniques, using instruments that are highly portable and non-disruptive. This enables the observation of individuals from a wide range of populations for longer periods of time in their natural environment and in interactive communicative situations. Being able to use these techniques with language stimuli opens up many possibilities for assessing (implicit) language knowledge across a range of populations and settings; it also enables us to open up and explore new dimensions of the intricate relationship between physiology and cognition. Heart Rate Variability (HRV), an ANS measure that has long been used as diagnostic (for a history see Berntson et al., 1997), is of particular interest. HRV measures the variability between successive heartbeats and captures the fluctuation of the length of the heartbeat intervals: low HRV is indicative of a highly regular heart rate, while high HRV reveals a highly irregular heart rate. The time between beats is measured in milliseconds (ms) and is called a regular-to-regular (R–R) interval or inter-beat interval (IBI). Changes in heart rate variability mainly capture the vagal innervation of the heart (Berntson et al., 1997) and HRV is considered an index of autonomic control of the heart: the variations are thought to result mainly from the dynamic interaction between the parasympathetic and the sympathetic inputs to the heart through the sinoatrial node (Forte, Favieri, & Casagrande, 2019; Thayer & Lane, 2000), although some argue that HRV only reliably indexes the cardiac activity of the parasympathetic nervous system (Malik, 1996; Reyes del Paso, Langewitz, Mulder, Van Roon, & Duschek, 2013; Laborde, Mosley, & Thayer, 2017; Thomas, Claassen, Becker, & Viljoen, 2019; Huang, Ko, & Liao, 2022). While the mechanics behind HRV responses require further investigation to be fully understood,[3] as is the case for the link between pupil dilation and effort, the applied literature generally accepts that HRV indexes stress: several studies have revealed a correlation between HRV variation and psychologically stressful situations whereby mental stress leads to an increase in interval regularity and thus a decrease in heart rate variability (for a meta-analysis of the use of HRV to diagnose stress, see Kim, Cheon, Bai, Lee, & Koo, 2018).

1.2. *Heart rate variability as an expression of autonomic nervous system activity and its relation to cognitive function*

Forte et al. (2019) systematically reviewed data on the link between resting-state HRV, as recorded by a continuous electrocardiogram in all but one case, and cognitive function from 19,431 healthy individuals. They found that higher HRV, in both the time and frequency domains (where high/medium/low frequency bands are distinguished), were associated with better cognitive performance, even after adjustment for the confounding variables commonly associated with HRV (i.e., age, gender, years of education, body mass index, blood pressure, cardiovascular diseases). This effect was observed across the domains of global cognitive functioning, attention, processing speed, executive functions, memory, language, and visuospatial skills. This conclusion highlights the important role of the ANS in cognitive functioning and confirms that ANS measurements can be used as a proxy for specific aspects of internal information processing that are not amenable to conscious retrieval and articulation.

However, Forte et al. (2019) also pointed out that ANS measurements remain vanishingly rare in the study of language. Among the exceptions are Britton et al. (2008) who tested short-term verbal memory, reasoning (Alice Heim 4-I), vocabulary, phonemic and semantic fluency in 5375 middle-aged adults. Their data includes the Mill Hill vocabulary test where the task is to explain the meanings of words or (in an alternative form of presentation) to select the correct synonym for each word from a list of six alternatives provided. They did not find any consistent relationship between HRV and language cognition, using both frequency- and time-related measures, on any of the tests. The study run by Zeki Al Hazzouri, Elfassy, Carnethon, Lloyd-Jones, and Yaffe (2017) tested participants at two points in time. At first contact, they took 10-s 12-lead electrocardiogram recordings from 2118 middle-aged adults and calculated two measures of HRV, standard deviation of normal-to-normal intervals (SDNN) and the root mean square of successive differences (RMSSD). Five years later, 3 cognitive tests were administered for verbal memory (the Rey Auditory-Verbal Learning Test), processing speed (the Digit Symbol Substitution Test), and executive function (the Stroop interference task). It was found that lower SDNN is associated with worse executive function only (as measured by Stroop) among middle-aged adults, above and beyond cardiovascular

[3] The models that have been developed to explain this link, such as the Neurovisceral Integration Model (Thayer, Hansen, Saus-Rose, & Johnsen, 2009), propose that HRV is linked to prefrontal cortex activity via the vagus nerve, which connects the heart and the brain. Individual differences in vagally mediated HRV translate into differences in prefrontal cortex activity and HRV, an index of cardiac vagal tone, has indeed been found to predict performance on several cognitive control tasks that rely on the prefrontal cortex, a key area that drives cognitive control (Colzato, Jongkees, de Wit, van der Molen, & Steenbergen, 2018): lower HRV is indicative of poorer cognitive function.

risk factors.

Frewen, Finucane, Savva, and al. (2013) included the language subtests on the Montreal Cognitive Assessment, i.e., language via confrontation naming with low-familiarity animals and repetition of complex sentences, when testing a sample of 4763 older adults (mean age 61.7). They likewise used both frequency- and time-related measures but did report that reduced HRV is associated with lower linguistic performance, even after adjustment for confounding variables of demographic, clinical and behavioural nature. Mahinrad et al. (2016) studied 3583 older participants with a mean age of 75 years. They recorded baseline 10-s ECGs from which the SDNN was calculated as index of HRV. Four cognitive domains were tested: selective attention, reaction time, processing speed and memory (using the picture-word learning test). Individuals were assessed at baseline and again during a mean follow-up period of 3.2 years. Lower HRV at baseline was associated with worse performance in reaction time (as measured by performance on the Stroop task) and processing speed (as measured by the Letter-Digit coding test). During follow-up, participants with lower HRV had a steeper decline in processing speed.

Given this, ANS measures such as HRV might well be useful in detecting linguistic knowledge that individuals have. However, although autonomic nervous activity has been shown to be strongly related to numerous cognitive functions, there is a paucity of research looking at language specifically in relation to cardiovascular expressions of ANS activity. The goal of this paper is, therefore, to test, for the first time, whether a cardiovascular response can help us detect grammatical knowledge in individuals and to explore how sensitive the cardiovascular response is to grammatical violations.

2. Methods

The current study goes beyond previous work in four crucial respects. Firstly, existing studies, surveyed above, that have included cognition, have looked at the relationship between reduced HRV and measures of cognitive function for language, rather than between HRV and linguistic knowledge per se. Secondly, if linguistic stimuli were used, these were virtually exclusively related to the lexicon in that the tasks were intended to assess memory for vocabulary. Yet in order to use language proficiently, a thorough command of grammar is required. Third, in existing studies, resting state HRV was correlated with cognitive function for language; in this study, we focus on acute HRV reactivity in response to a particular stimulus. And lastly, previous studies have contrasted the presence with the absence of a property, but have not tested sensitivity to different levels of intensity, which would be a prerequisite for research in the area of language, but also applies to other areas of cognition where the gradedness of the stimulus is important (e.g., general goal difficulty in Gellatly & Meyer, 1992; or the nature of reward vs. punishment in Löw, Lang, Smith, & Bradley, 2008; Gu, Bai, & Wang, 2015). As a starting point, we focus on individuals' overall performance in an error detection task that is presented as a speech rating task, where stimuli are manipulated for the density of grammatical errors against the background of accent. If HRV responses are sensitive enough to reflect the recognition of grammatical violations, HRV measures should be able to differentiate the speech stimuli depending on the manipulated conditions.

2.1. Participants

A total of 41 healthy adults who were native speakers of British English (21 female) were recruited and invited to the lab through social media such as Facebook. Their age ranged from 18 to 44 ($M = 21.9$, $SD = 6.4$), and none of them were diagnosed with any learning difficulty or cardiac rhythm issues. Twelve of the participants held a degree of higher education. 27 out of 40 participants spoke another language (with 10 reporting a native-like level in their second language, and one speaking a Slavic language). All of them reported to be familiar with foreign accented English speech to some degree (with at least monthly contact with non-native speakers).

2.2. Speech rating task

Participants were asked to listen to 40 short English speech samples, half of which contained grammatical errors against articles (e.g., a/an, the). To examine the sensitivity of physiological responses to grammatical errors in different contexts, the speech samples were presented in different accents (native vs. foreign) and by speakers of different gender (female vs. male), as these conditions have been found to affect listeners' responses (Hanulíková, Van Alphen, Van Goch, & Weber, 2012; Linek, Gerjets, & Scheiter, 2010).

Each speech sample was created based on a transcribed response to an interview question about one of eight common topics (i.e., education, environment, culture, globalisation, health, city, minority group, addiction). The transcripts were extracted from the BACKBONE[4] English as Lingua Franca (ELF) Corpus of Polish speakers (i.e., the original interviewees). Firstly, 40 samples of transcribed responses (5 samples per topic) were selected and examined by a qualified linguist and by a trainee linguist (a native speaker of British English) to correct any language errors other than errors against articles. Then, half of the samples were designated to represent the error condition; in these, additional English article errors were inserted to reach the desired error density levels.

English article errors were used as they are frequent in the English speech of non-native speakers. In English, one in five words is estimated to be an article (Dryer, 1989) but linguists have not been able to provide a concise and comprehensive description of how the article system works (Yoon, 1993). Moreover, these errors are not usually corrected as they do not tend to impede communication

[4] The corpus was developed via the EU project "BACKBONE – Corpora for Content & Language Integrated Learning", which was funded by the EU Lifelong Learning Programme, 143502-LLP-1-2008-1-DE-KA2-KA2MP, 2008–2010.

(Vann, Meyer, & Lorenz, 1984). As shown in Example A below (with article errors and their nouns underlined), some articles were redundant (*a globalisation*) while some were omitted (*same music/books*); some definite and indefinite articles are substituted (*the positive or negative impact*). The context provided by the coherent discourses across samples on the same topic allows for different types of article errors (e.g., misuse of definite articles).

Example A. with errors

> I think that culture is one of the areas most affected by a globalisation and it's hard to say whether it is the positive or negative impact. I think that thanks to a globalisation, people all around the world listen to same music, watch the same movies, and read same books. They can discuss the same issues with each other, and understand each other better, because they know what they are talking about.

Example B. without errors

> I think that immigration is a positive outcome of globalisation because people have the possibility to meet each other, to travel from one country to another and to get a job in a new country. This helps with integration and creating a multicultural environment. However, some emigrants live in very bad conditions and sometimes they are unwilling to integrate with the native citizens of the country where they live.

The density of grammatical errors, calculated as the number of errors divided by the number of nouns, varied from 18% to 56% among the 20 samples with errors. For example, the sample with the lowest error density (18%) comprised five medium-to-long sentences (92 words in total) and had only four errors (no more than one error in each sentence). On the contrary, the speech sample with the highest error density (56%) comprised four shorter sentences (63 words in total) but five errors.

Based on the 40 written samples, 160 speech samples were recorded by four speakers who were either native or foreign English speakers (British and Polish, female and male). The length of the speech samples varied from 13 to 39 s ($M = 25$). To counterbalance the combination of content and voice, four sets of experimental stimuli were created using a Latin square design (set A, B, C, and D as shown in Table 1 in SupMatA) and allocated randomly to participants. Each participant listened to eight blocks, with two blocks in the voice of the same speaker and each block consisting of five speech samples about the same topic. Each block was independently rated by a different group of 60 native speakers ($M_{age} = 36.5$, 30 female) for intelligibility (extent of understanding) and comprehensibility (ease of understanding) on a scale from 1 to 100. The majority of the speech samples were rated as highly intelligible and comprehensible (see Table 2 in SupMatA for the ratings distribution).

During the task, participants were seated comfortably in front of a computer in a quiet lab. They were given at least 10 minutes to return to normal and slow breathing between arriving at the lab and before starting the task. Speech stimuli and instructions were presented to them through BioGraph Infiniti software by Thought Technology. Participants were instructed to listen to the audio samples and were asked to rate on an iPad how much they would like to be represented by each speaker in terms of argument and language (note that each participant listened to all stimuli and hence heard all arguments). After every block of five speech samples, participants gave two ratings on a 7-point Likert scale ranging from "not at all" on the left end (−3) to "very much" on the right end (3). During the speech rating task, participants took a short break of 5 s after each trial, and a longer break of 90 s after four blocks of trials while listening to relaxing music (i.e., resting state). The session took about 30 minutes in total.

Thought Technology's 5-channel ProComp 5 encoder was used to track participants' cardiovascular activity. A BVP-Flex/Pro sensor was attached to the middle finger of their non-dominant hand, to record the blood volume pulse (BVP) signal (in millivolts) at a rate of 2048 samples per second. Participants were asked to find a comfortable position for their non-dominant hand and keep it still till the end of the task to avoid artifacts. Cardiovascular measures were calculated for each trial, while the baseline measures (i.e., resting state) were based on the 90 s recording during the longer break. We also collected skin conductance data as supplementary automatic nervous system measure. As the focus of the current study is on heart rate variability, the electrodermal measures are reported in SupMatC.

Upon completion of the task, the participants were given an exit survey. Recall that all speakers listened to samples by all four speakers and that article errors were the only errors present in the stimuli. Participants were first asked about the origins of the foreign accent; 5 did not identify the accent as Polish. They were also asked whether there were any errors in the speech they heard, to which only 1 participant responded negatively. When asked to estimate the number of errors they heard from each speaker on a scale from 0 to 100, participants showed substantial variability with the British female voice being given a much lower error estimate than the other speakers ($M = 16.54$ and $SD = 18.49$ for British female speaker, $M = 23.73$ and $SD = 25.35$ for British male speaker, $M = 22.20$ and $SD = 18.47$ for Polish female speaker, $M = 22.83$ and $SD = 18.67$ for Polish male speaker). Participants were then asked to describe the type of errors they heard in their own words. While no single participant uniquely identified article errors, only 7 participants did not name article errors at all; many were unaware of the name of the category, declaring instead that "the letter 'a' was missing" or that 'the' was "forgotten or added in unnecessary places". When asked to tick "all that applied" from a list of 6 errors and suggest any other errors they heard, only 2 participants selected/suggested "incorrect use of articles" as only category alongside "incorrect pronunciation" and 4 did not include "incorrect use of articles" at all.

2.3. Cardiovascular measures of ANS activity

The cardiovascular measures based on blood volume pulse data were recorded and extracted using BioGraph Infiniti software (for details see SupMatB). In addition to heart rate variability measures, heart rate and pulse amplitude were included as supplementary

cardiovascular measures: heart rate variability measures reflect parasympathetic nervous system activity (Laborde et al., 2017; Malik, 1996); heart rate and pulse amplitude measures reflect both parasympathetic and sympathetic nervous system activity (Kantono et al., 2019; Salimpoor, Benovoy, Longo, Cooperstock, & Zatorre, 2009; Lin, Lin, Lin, & Huang, 2011; Soni & Rawal, 2020.

Heart rate variability analysis can be conducted in the time domain, in the frequency domain (where high/medium/low frequency bands are distinguished), and by using non-linear analyses (Forte et al., 2019). No frequency-domain HRV measures were exported because a minimum 64 s of recording is required to produce a value by the software, yet none of our trials were longer than 39 s. HRV analysis can also be conducted in the time domain and two time-domain HRV measures, NN50 and RMSSD, were exported via the session statistics report from BioGraph Infiniti. Line graphs were inspected visually for artifacts via the session reviewing function, i.e., sudden dips or spikes in the data caused by movement of the sensor or finger which disturb the regular patterns of finger pulse data. Trials with artifacts (34 out of 1640) were removed from subsequent data analysis.

Both exported HRV measures reflect beat-to-beat variance via the differences between successive normal sinus (NN) intervals. More specifically, NN50 refers to the number of pairs of adjacent NN intervals differing by more than 50 ms in each recording session, while RMSSD captures the root mean square of differences between adjacent NN intervals. Rather than using raw HRV measurements, we opted for relative values (see also Christensen, 2012) and calculated the difference between the value measured during the performance on the task and the resting state (i.e., during the 90s break in the middle of the task). This was done for each measure, for each participant and each item. In that sense, we obtained a measure of (relative) change to the relaxed state, when a participant was not engaged in any task.

Although longer recording sessions (>5min) are preferred conventionally, ultra-short-term periods (<1min) have been found to predict long-term measures reliably (e.g., Baek, Cho, Cho, & Woo, 2015; Salahuddin, Cho, Jeong, & Kim, 2007). For example, Salahuddin et al. (2007) validated that mental stress in mobile settings can be reflected by RMSSD from a 30-s recording, where both measures showed significant differences between baseline and Stroop test conditions. Furthermore, in our design, many similar events were used (half of the stimuli contained errors, while the other half did not contain errors), and the data for all error-ridden stimuli was considered as being of the same type, as was the data for all stimuli that were grammatically correct. Capturing a particular physiological measure at a particular point in time is a unique event and using many similar events helps obtain more stable measurements. Overall, participants heard from 503 to 510 s ($M = 507.5$) of grammatically correct recordings, and from 502 to 509 s ($M = 505.3$) of grammatically incorrect recordings.

3. Data analysis and results

Faced with a literature that lacks firm agreement as to the type of ANS activity the different measures reflect and relies on a wide range of variables to capture HRV, we turned to Graphical Modelling[5] (cf., Lauritzen, 1996; Edwards, 2000; Højsgaard, Edwards, & Lauritzen, 2012) to establish the direct and indirect relationships within the variables typically used in the literature on physiological measures and language cognition.

3.1. Variable selection

Simplified, a graphical model represents the variables of interest (vertices) and their relations (edges), where missing edges indicate conditional independence (CI; cf., Németh & Rudas, 2013). The approach aims to find CIs in a set of correlated variables in order to simplify the joint density of these variables as much as possible. More formally, if X_A, X_B, and X_C are such variables, we establish a conditional independence of X_A and X_B given X_C if:

$$f(X_A, X_B|X_C) = f(X_A|X_C)f(X_B|X_C)$$

where $f(\,)$ denotes probability mass function. This statistical procedure can also reveal variables that are *endogenous*, meaning they are *determined* (i.e., informed) by their relationships with other variables within the model. For example, a variable such as word orthographic density can be informed by other variables, like word length and frequency, bigram frequency and others, but at the same time, it does not inform any other variable itself. Similarly, an *exogeneous* variable would be informative about other variables in the model, while none of the remaining variables would provide information about that *exogeneous* variable.

We applied Graphical Modelling specifically to determine if our set of physiological measures contains an endogenous variable that would be used in further analyses. The logic of this approach is simple: as endogenous variables do not inform about but are informed by other variables in the given set, they can be used as representative or proxy of that set of variables. In other words, rather than using a principal component (i.e., a composite variable), we searched for the variable that is informed (or determined) by other skin-

[5] The standard statistical approach in a case that offers multiple independent variables would be to look for indicators that are highly correlated as such a relationship would suggest that these variables might share a *common source*. This new, latent variable is typically isolated as an optimally weighted composite (i.e., Principal or Independent Component), which can enter into further analyses. The problem with this approach is that it can result in something that is difficult to understand and interpret. Further to this, in our case, the great majority of possible bi-variate correlations was low ($r < 0.2$), with two exceptions only: moderate $r = 0.53$ between SC.SD and SC.M ($t = 25.04, p < 0.0001$), and low $r = 0.28$ between HRV.NN50 and HR.SD ($t = 11.74, p < 0.0001$). This might not be surprising as such co-relationships between the position (mean) and scale (standard deviation) parameters have been found for other response measures too (e.g., for response times: Wagenmakers & Brown, 2007; for heart period variability and mean period length: Fleiss, Bigger, & Rolnitzky, 1992).

conductance measures. Similar to running a Principal or Independent Component Analysis, this proposed approach is data-driven and replicable, and, at the same time, more parsimonious than the alternative of analysing all possible dependent variables, which posits a problem of multiple tests and, hence, increases the chance of "discovery" (i.e., null-rejection).

To run a Graphical Model we utilised the **pcalg** package (Kalisch, Maechler, Colombo, Maathuis, & Buehlmann, 2012) in the **R** software environment (Team, 2022). For visualization we made use of the **igraph** (Csardi & Nepusz, 2006, p. 1695), and **gRim** (Højsgaard et al., 2012) packages for **R**. Given our focus on cardiovascular measures (Fig. 1), in addition to the HRV measures of interest (HRV.NN50 and HRV.RMSSD), we also included several other available heart rate measures (HR.M, HR.SD, PULSE.M, PULSE.SD) used in the literature to allow for competition in the identification of the best endogenous candidate. Although, as explained above, all variable values were relative to the resting state, they were still rather different in location and scale, which means not directly comparable. To facilitate statistical modelling and comparability we transformed all difference-variables using a rank-normal (Gaussian) transformation, converting differences to standardised units (i.e., z-values).

Fig. 1 shows that, within the set of cardiovascular measures, HRV.NN50 and HR.M are both candidate endogenous variables. In fact, they are determined by other variables, either directly (PULSE.M, HR.SD, HR.M, and HRV.RMSSD) or indirectly (PULSE.SD). Finally, HR.M and HRV.NN50 are mutually co-dependent (as represented by a direct connection, i.e., a two-way arrow, in Fig. 1). Thus, within the set of cardiovascular measures, HR.M and HRV.NN50 are the best candidate-representatives for the set of biofeedback measures. The main model considered HRV.NN50 as the dependent variable and HR.M as its covariate (i.e., continuous predictor), since, in statistical terms, variability depends on the mean tendency. The model with HR.M as dependent variable and HRV.NN50 as its covariate are included in SupMatC, for completeness.

3.2. Statistical model

The complete dataset available for modelling consisted of $N = 1640$ datapoints. As the focus of the study is on testing whether manipulation of the linguistic variable ErrorDensity has an observable effect on the dependent variable HRV.NN50, we removed all measurements of experimental items where there were no errors (ErrorDensity $== 0$), leaving a dataset to $N = 820$. This decision follows standard practice across the range of two-alternative forced choice tasks (2AFC); for example, in lexical decision data, typically only real word latencies are considered. The same applies to time latency analyses for semantic or grammaticality judgements and other similar tasks. Finally, we removed an additional 16 points as artifacts, retaining a final dataset of size $N = 804$.

Using the **mgcv** (Wood, 2006, 2011) and **itsadug** (van Rij, Wieling, Baayen, & van Rijn, 2022) packages for **R**, we applied a Generalised Additive Mixed Effect Model, which allowed us to test for a possible non-linear relationship between ErrorDensity and the main dependent variable HRV.NN50, as well as to specify random factor smooths of Participants over experimental Trials. Random factor smooths allow for an efficient account of the individual random variation over the course of the experiment (which can be due to, e.g., attention loss, distraction, fatigue etc.). Our base model was specified as follows:

$$gam(\textbf{HRV.NN50} \sim \text{SpeakerAccent} + \text{s}(\text{TrialOrder}) + s(\text{ErrorDensity}) + s(\text{HR.M}) + s(\text{TrialOrder}, \text{Participant}, bs='\text{fs}', m = 1), \ldots)$$

The model has three fixed effect terms: a 2-level factor (SpeakerAccent: British vs. Polish) and two smoothed covariates, i.e., TrialOrder that was scaled (z-transformed) to control for its typically large variance, and ErrorDensity. As explained above, the dependent variable HRV.NN50 was rank-normally transformed.

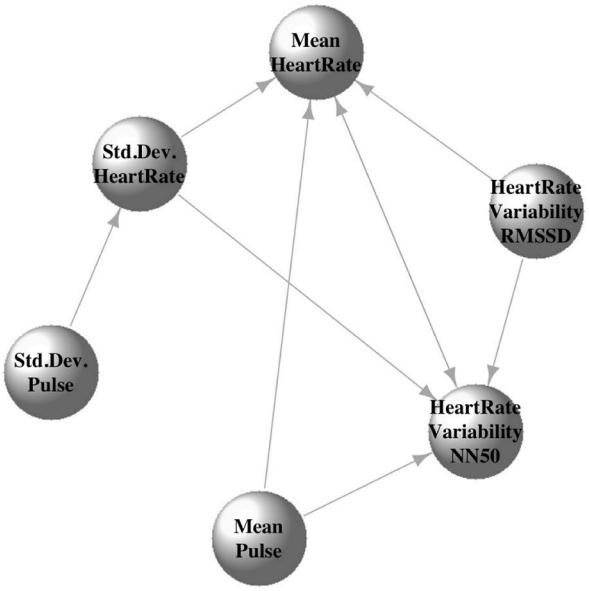


Fig. 1. Results of graphical modelling for cardiovascular measures.

Table 1
Trimmed (i.e., criticised) Generalised Additive Mixed Effect Model fitted to heart rate variability (HRV.NN50), on the subsample of the data containing errors.

Parametric coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	−0.047	0.146	−0.320	0.749
SpeakerAccent (Polish)	0.047	0.014	3.255	0.001
Smooth terms:				
	edf	**Ref.df**	**F**	**p-value**
s(TrialOrder)	1.000	1.000	0.926	0.336
s(ErrorDensity)	2.863	3.465	12.739	<0.0001
s(HR.M)	4.164	5.202	13.555	<0.0001
s(TrialOrder, Participant)	81.711	368.000	45.494	<0.0001

3.2.1. *Modelling heart rate variability (HRV.NN50)*

We applied model criticism to detect if the effects are affected by influential datapoints, i.e., outliers or extremes, where the former can deflate and the latter may inflate the magnitude of an effect (Baayen & Milin, 2010). Only 15 datapoints fell outside the interval of 2.5 standardised residuals and were removed. The trimmed model showed an increase in significance of critical predictors: SpeakerAccent, ErrorDensity, and HR.M. Overall, the model fit is very strong: $R^2_{Adjusted} = 0.96$; Deviance explained = 96.3% (for the untrimmed model: $R^2_{Adjusted} = 0.94$; Deviance explained = 95%). The trimmed model is summarised in Table 1 and its main effects of SpeakerAccent, ErrorDensity, and HR.M are presented in Fig. 2.

The effect of TrialOrder appeared linear (edf and Ref.df are 1.0) and not significant. It was, however, retained among the model's fixed effects as it figures in the random part as well, revealing the participants' non-linear trajectories over the experimental trials.

As hypothesized, there was a significant effect of ErrorDensity on HRV.NN50: the mid panel in Fig. 2 shows a steep decrease in Heart Rate Variability when the ErrorDensity is between 0.2 and 0.4. From about 0.45, the confidence intervals include zero, indicating that the effect should not be considered in this region; moreover, when the number of knots is reduced to 3, the curve attenuates. In simple terms, if speech is punctuated by errors, Heart Rate Variability decreases at a rate that is in line with the density of the errors, up to a density of 0.4.

SpeakerAccent also influences HRV.NN50: HRV.NN50 is lower, i.e., less variable, when errors are made by native speakers of English. This is visualised in the left panel in Fig. 2. The effect is rather mild, as indicated by the relatively wide Confidence Intervals, but as the strength of the effect increased after trimming (p-value reduced from 0.016 to 0.001), the effect can be considered reliable.

Average heart rate, which was included as a control predictor, also revealed the expected relationship: the higher the mean heart rate, the lower the heart rate variability (right panel in Fig. 2).

As an additional test, we ran a comparable model on data for correct items, which implies that, if compared with the original model specification above, the term s(ErrorDensity) had to be removed. Crucially, the effect of SpeakerAccent remained significant (p-value = 0.03; and 0.01 for the untrimmed model).

4. Discussion

Over the past decades, focus has been on developing methods that allow tapping into cognition, and in particular into those aspects of cognition that are not directly observable. Linguistic knowledge and skills are an area that is in particular need of such implicit methods, as linguistic knowledge develops largely without awareness and is deployed without awareness which may make it difficult or impossible to articulate without formal linguistic training. Yet, accurately assessing an individual's linguistic abilities is important for a number of reasons, including designing interventions that prevent or address delays in cognitive development or deterioration of brain health.

The goal of this paper was to test, for the first time, whether a cardiovascular response such as HRV that taps into the parasympathetic system can help us detect whether individuals possess grammatical knowledge and explore how sensitive the cardiovascular response is to linguistic stimuli. Up until now, pupil size, which reflects the workings of the sympathetic nervous system, has been the ANS measure of choice for detecting language knowledge without relying on explicit judgments. In our study, we exposed 41 healthy native speakers of British English to audio-recordings of speech stimuli that were either grammatically correct or contained errors against the English article system and differed in the density of these errors. The texts were read in a native British and a Polish accent by both a male and a female voice. Participants were instructed to listen to the four different speakers in both error-free and error-ridden conditions and were occasionally asked to rate how much they would like to be represented by each speaker.

Taking into account the properties of the dataset, the data was analysed using a Generalised Additive Mixed Effects Model. The model confirmed that there is a cardiovascular response to grammatical violations: we registered a statistically significant reduction in HRV as indexed by NN50 in response to stimuli that contain errors. More specifically, the cardiovascular response reflects the extent of the linguistic violations, and HRV decreases in line with an increase in the number of errors. Listening to speech containing errors reduces HRV, and the reduction increases with the number of violations, up to a certain level. Recall that all participants heard and rated all stimuli, hence the degree to which participants agree with the arguments in the stimuli does not affect our findings. The fact

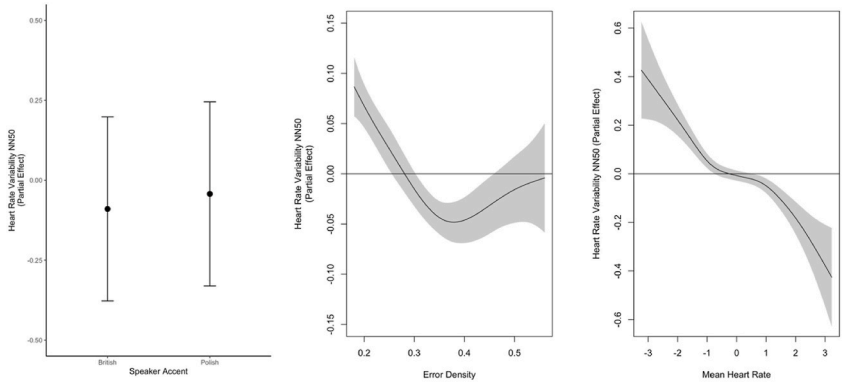


Fig. 2. Effect of speaker accent (left panel), error density in the sample of items (mid panel), and average heart rate (right panel) on HRV as indexed by NN50.

that there is a cardiovascular response to the violation of regularities that can be observed in language, and that the cardiovascular response becomes stronger as the violations become more frequent, supports the assumption that language users have absorbed the usage patterns that are typical of their language, have come to expect them, and therefore respond to them being violated. HRV measures can thus be used to reveal linguistic knowledge that language users have, without relying on their ability to *declare* their knowledge. The observation that linguistic knowledge can be detected using cardiovascular measures brings into focus a new dimension of the intricate relationship between physiology and cognition and opens up new pathways for exploring this link.

The knowledge language users have is not limited to grammatical linguistic information: their expectations go beyond the form, into the social domain. In addition to sensitivity to error density, participants were also sensitive to accent, with a British accent triggering a slight but significant reduction in HRV in sentences containing grammatical violations compared to a Polish accent. In other words, hearing speakers with a native accent commit errors against a core property of their native language reduces HRV more than hearing speakers with a foreign accent make those same errors. A foreign accent is the hallmark of the speech of someone who has learned the language at a later age (Moyer, 2013), as very few later learners manage to develop a native-like pronunciation. Later learners are also known to struggle with (aspects of) grammar in their second language (Abrahamsson, 2012), in particular with grammatical structures that are absent from their first language (Sabourin, Stowe, & De Haan, 2006). Taken together, these two facts make errors more expected in foreign-accented speech than in native-accented speech, which explains why a Polish accent triggered less of a reduction in HRV, and a British accent triggered more of a reduction.

What does the reduction in HRV as indexed by NN50 in response to stimuli that contain errors signal? To answer this question, we combine insights gleaned from both the error and accent manipulations of our study. Recall that the majority of the speech samples were rated as highly intelligible and comprehensible; therefore, the HRV response to violations is unlikely to index difficulty of understanding. This conclusion is further supported by the observation that identical errors trigger a different response depending on accent, and that the response is stronger for the more familiar accent; errors in native accented highly intelligible speech would not be expected to increase difficulty of understanding more than errors in its foreign accented equivalent. A more plausible interpretation would suggest that changes in HRV as indexed by NN50 are to be interpreted in terms of expectation violation, in the first instance: errors are unexpected, and more so in native accented speech. Expectation violations tend to require more cognitive effort, thereby putting stress on the cognitive system, and this is picked up by the physiological system. Recall that the applied literature generally accepts that HRV indexes stress: mental stress leads to an increase in interval regularity and thus a decrease in heart rate variability (for a meta-analysis of the use of HRV to diagnose stress, see Kim et al., 2018).

Our HRV-based results are in line with findings from pupillometry on the effects of complexity, conflict, accent and errors on language processing. Several early pupillometric studies on language processing (Ben-Nun, 1986; Just & Carpenter, 1993) found that pupil size increased when more complex information had to be processed; the pupillary response was therefore considered as an indicator of how intensely the processing system is operating. The same effect was found in the auditory processing of temporary ambiguities, a.k.a. the garden path effect, in the presence of congruent or incongruent prosodic cues: pupil diameter reliably increased in case of conflicts between syntax and prosodic cues. Studies have also looked at the effect of accent on pupil size. Porretta and Tucker (2019) focused on the way in which the pupil responds to foreign-accent related intelligibility and found a negative correlation between the size of the pupil and the intelligibility of the sentence: as intelligibility decreases, pupil dilation increases. Likewise, they observed individual differences: listeners with more experience interacting with accented speech displayed reduced dilation overall, but high experience listeners had a higher threshold at which reduced intelligibility elicited greater dilation. As far as accent is concerned, Brown, McLaughlin, Strand, and Van Engen (2020) found, again using pupillometry, that even fully intelligible accented speech triggers an increase in pupil diameter which was taken to signal a larger burden on the cognitive system: resolving deviations between the acoustic input and stored lexical representations incurs a processing cost. However, they also reported that listeners habituate quickly to the accent: after 20 sentences (which each were on average 2*s* long), listeners appeared to have accommodated the accent. There is evidence from pupillometry that errors likewise increase cognitive load. Hubert Lyall and Järvikivi (2021) investigated a range of errors or violations of morpho-syntactic, semantic and socio-cultural expectations and recorded an increase in pupil size in

response to an error or anomaly.

We started from the claim that linguistic knowledge, which is typically acquired and deployed implicitly, benefits from being measured in an implicit fashion. Linguists generally agree that the knowledge language users have of their first language is implicit: except for the points that are discussed in educational settings, most language users are able to produce language correctly yet are unable to reference the rules that guide usage. In the same spirit, measures are considered implicit if they are taken while language users do something they naturally do, e.g., listen to a recording or read a text, and are not aware of what is being measured. In our study, participants listened to audio recordings and were asked, after every block of five speech samples, how much they would like to be represented by each speaker in terms of argument and language. Data from the exit interview suggests that, when asked, participants did report awareness of errors. This explicit *noticing* of errors might also render the knowledge *assessment* explicit. However, relying on the protocol used in studies of implicit knowledge, we established that it remained beyond the participants' ability to uniquely identify the article errors; this supports the idea that the *knowledge* itself was indeed implicit and suggests that he *assessment* of this knowledge likewise remained implicit, i.e., that participants were not aware we were testing their knowledge of the English article system specifically.

5. Conclusions

Implicit measures infer mental contents from responses on performance-based tasks, thereby enabling capturing knowledge, thoughts and feelings that people are either unwilling or unable to report. Physiological measures are often used where access is needed to those aspects of cognition that are not directly observable or where there is a need to circumvent the requirement to provide an explicit evaluative judgment, e.g., in work with populations that cannot yet or can no longer express their opinion, due to (young/old) age or ill health (be it physical or cognitive/mental). Building on the relation between language cognition and the nervous system, we have provided first evidence to suggest that Heart Rate Variability, a cardiovascular measure of autonomic nervous system activity, can be used as indicator of implicit linguistic knowledge. Departures from linguistic normality trigger a clear cardiovascular reaction, and thereby reveal linguistic knowledge on the part of the individual without the need for explicit articulation. This observation brings into focus a new dimension of the intricate relationship between physiology and cognition, suggesting that cognitive effort reverberates through the physiological system in more ways than previously thought. Because HRV recordings can be made using portable and non-intrusive systems this approach offers possibilities for use in natural environments and with a wider range of populations than the standardly used instruments do. This, in turn, improves the ecological validity and representativity of the findings and makes the approach suitable for research on language in both clinical and non-clinical settings.

Funding

This work was funded by a Leverhulme Trust Research Leadership award (RL-2016-001) to Dagmar Divjak which funded all authors.

Open science statement

Materials, data and code are available from our anonymous OSF repository https://osf.io/jn24s/?view_only=01ecbb1c63b84195bf2994cc1a441a14.

Ethics

The study was approved by the Ethics Committee of the University of Birmingham, UK.

Author statement

Dagmar Divjak: Conceptualization, Formal analysis, Funding acquisition, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing. Petar Milin: Conceptualization, Data curation, Formal analysis, Methodology, Visualization, Writing – original draft (Results section). Hui Sun Investigation, Methodology, Writing – original draft (Methods section), completed while the author was at Cardiff.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that support the findings reported here are openly available on the University of Birmingham Institutional Research Archive, UBIRA at https://edata.bham.ac.uk/1013/. The R-scripts for the analyses are available on the GitHub repository at https://github.com/ooominds/heart-rate-variability-as-an-indicator-of-language-knowledge.

Supplementary Materials

Supplementary materials to this article can be found online at https://doi.org/10.1016/j.jneuroling.2023.101177.

References

Abrahamsson, N. (2012). Age of onset and nativelike L2 ultimate attainment of morphosyntactic and phonetic intuition. *Studies in Second Language Acquisition, 34*(2), 187–214.

Ambach, W., & Gamer, M. (2018). Chapter 1 - physiological measures in the detection of deception and concealed informatio. In J. P. Rosenfeld (Ed.), *Detecting concealed information and deception* (pp. 3–33). Academic Press.

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research, 3*(2), 12–28.

Baek, H. J., Cho, C.-H., Cho, J., & Woo, J.-M. (2015). Reliability of ultra-short-term analysis as a surrogate of standard 5-min analysis of heart rate variability. *Telemedicine and e-Health, 21*(5), 404–414. https://doi.org/10.1089/tmj.2014.0104

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin, 91*, 276–292. https://doi.org/10.1037/0033-2909.91.2.276

Beatty, J., & Lucero-Wagoner, B. (2000). *The pupillary system*.

Ben-Nun, Y. (1986). The use of pupillometry in the study of on-line verbal processing: Evidence for depths of processing. *Brain and Language, 28*(1), 1–11.

Berntson, G. G., Bigger, J. T. J., Eckberg, D. L., Grossman, P., Kaufmann, P. G., Malik, M., et al. (1997). Heart rate variability: Origins, methods and interpretative caveats. *Psychophysiology, 34*, 623–648.

Britton, A., Singh-Manoux, A., Hnatkova, K., Malik, M., Marmot, M. G., & Shipley, M. (2008). The association between heart rate variability and cognitive impairment in middle-aged men and women. The Whitehall II cohort study. *Neuroepidemiology, 31*(2), 115–121. https://doi.org/10.1159/000148257

Brown, V. A., McLaughlin, D. J., Strand, J. F., & Van Engen, K. J. (2020). Rapid adaptation to fully intelligible nonnative-accented speech reduces listening effort. *Quarterly Journal of Experimental Psychology, 73*(9), 1431–1443.

Chapman, L. R., & Hallowell, B. (2015). A novel pupillometric method for indexing word difficulty in individuals with and without aphasia. *Journal of Speech, Language, and Hearing Research, 58*(5), 1508–1520.

Christensen, S. C. (2012). Working memory in adults with aphasia: Considering effort invested through a physiological measure - heart rate variability. (Ph.D. Speech and hearing science), Arizona state. Retrieved from https://keep.lib.asu.edu/items/151047.

Colzato, L. S., Jongkees, B. J., de Wit, M., van der Molen, M. J., & Steenbergen, L. (2018). Variable heart rate and a flexible mind: Higher resting-state heart rate variability predicts better task-switching. *Cognitive, Affective, & Behavioral Neuroscience, 18*(4), 730–738.

Csardi, G., & Nepusz, T. (2006). *The igraph software package for complex network research*. Complex Systems: InterJournal.

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin, 135*(3), 347.

Dryer, M. (1989). Article-noun order. In *Proceedings of the 25th regional meeting of the Chicago linguistic society* (pp. 83–97).

Edwards, D. (2000). *Introduction to graphical modelling*. Springer Science & Business Media.

Ellis, N. C. (2015). Implicit and explicit learning: Their dynamic interface and complexity. In P. Rebushat (Ed.), *Implicit and explicit learning of languages* (pp. 3–23). Amsterdam: John Benjamins.

Fleiss, J. L., Bigger, J. T., Jr., & Rolnitzky, L. M. (1992). The correlation between heart period variability and mean period length. *Statistics in Medicine, 11*(1), 125–129. https://doi.org/10.1002/sim.4780110111

Forte, G., Favieri, F., & Casagrande, M. (2019). Heart rate variability and cognitive function: A systematic review. *Frontiers in Neuroscience*. https://doi.org/10.3389/fnins.2019.00710

Frewen, J., Finucane, C., Savva, G. M., & al, e. (2013). Cognitive function is associated with impaired heart rate variability in ageing adults: The Irish longitudinal study on ageing wave one results. *Clinical Autonomic Research, 23*, 313–323. https://doi.org/10.1007/s10286-013-0214-x

Gawronski, B., & De Houwer, J. (2014). *Implicit measures in social and personality psychology*.

Gawronski, B., & Hahn, A. (2018). Implicit measures: Procedures, use, and interpretation. In *Measurement in social psychology* (pp. 29–55). Routledge.

Gellatly, I. R., & Meyer, J. P. (1992). The effects of goal difficulty on physiological arousal, cognition, and task performance. *Journal of Applied Psychology, 77*(5), 694.

Gu, L., Bai, X., & Wang, Q. (2015). Impact of reward/punishment conditions on behavioral inhibition and automatic physiological responses in the stages. *Acta Psychology Sinica, 47*(1), 39.

Hanulíková, A., Van Alphen, P. M., Van Goch, M. M., & Weber, A. (2012). When one person's mistake is another's standard usage: The effect of foreign accent on syntactic processing. *Journal of Cognitive Neuroscience, 24*(4), 878–887.

Højsgaard, S., Edwards, D., & Lauritzen, S. (2012). *Graphical models with R*. New York: Springer.

Huang, W.-L., Ko, L.-C., & Liao, S.-C. (2022). The association between heart rate variability and skin conductance: A correlation analysis in healthy individuals and patients with somatic symptom disorder comorbid with depression and anxiety. *Journal of International Medical Research, 50*(9), Article 03000605221127104.

Hubert Lyall, I., & Järvikivi, J. (2021). Listener's personality traits predict changes in pupil size during auditory language comprehension. *Scientific Reports, 11*(1), 1–15.

Just, M. A., & Carpenter, P. A. (1993). The intensity dimension of thought: Pupillometric indices of sentence processing. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 47*(2), 310.

Kalisch, M., Maechler, M., Colombo, D., Maathuis, M. H., & Buehlmann, P. (2012). Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software, 47*(11), 1–26.

Kantono, K., Hamid, N., Shepherd, D., Lin, Y. H. T., Skiredj, S., & Carr, B. T. (2019). Emotional and electrophysiological measures correlate to flavour perception in the presence of music. *Physiology & behavior, 199*, 154–164.

Kim, H.-G., Cheon, E.-J., Bai, D.-S., Lee, Y. H., & Koo, B.-H. (2018). Stress and heart rate variability: A meta-analysis and review of the literature. *Psychiatry investigation, 15*(3), 235. https://doi.org/10.30773/pi.2017.08.17

Laborde, S., Mosley, E., & Thayer, J. F. (2017). Heart rate variability and cardiac vagal tone in psychophysiological research–recommendations for experiment planning, data analysis, and data reporting. *Frontiers in Psychology, 8*, 213.

Lauritzen, S. L. (1996). *Graphical models* (Vol. 17). Clarendon Press.

Linek, S. B., Gerjets, P., & Scheiter, K. (2010). The speaker/gender effect: Does the speaker's gender matter when presenting auditory text in multimedia messages? *Instructional Science, 38*(5), 503–521.

Lin, H. P., Lin, H. Y., Lin, W. L., & Huang, A. C. W. (2011). Effects of stress, depression, and their interaction on heart rate, skin conductance, finger temperature, and respiratory rate: Sympathetic-parasympathetic hypothesis of stress and depression. *Journal of Clinical Psychology, 67*(10), 1080–1091.

Loewenfeld, I. E. (1993). *The pupil*. Ames, IA: Iowa State University Press.

Löw, A., Lang, P. J., Smith, J. C., & Bradley, M. M. (2008). Both predator and prey: Emotional arousal in threat and reward. *Psychological Science, 19*(9), 865–873.

Mahinrad, S., Jukema, J. W., van Heemst, D., Macfarlane, P. W., Clark, E. N., de Craen, A. J., et al. (2016). 10-Second heart rate variability and cognitive function in old age. *Neurology, 86*(12), 1120–1127. https://doi.org/10.1212/WNL.0000000000002499

Malik, M. (1996). Heart rate variability: Standards of measurement, physiological interpretation, and clinical use: Task force of the European society of cardiology and the north American society for pacing and electrophysiology. *Annals of Noninvasive Electrocardiology, 1*(2), 151–181.

Moyer, A. (2013). *Foreign accent: The phenomenon of non-native speech*. Cambridge: Cambridge University Press.

Németh, R., & Rudas, T. (2013). On the application of discrete marginal graphical models. *Sociological Methodology, 43*(1), 70–100.
Porretta, V., & Tucker, B. V. (2019). Eyes wide open: Pupillary response to a foreign accent varying in intelligibility. *Frontiers in Communication, 4*, 8.
Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372.
Rayner, K. (2009). The 35th Sir Frederick Bartlett Lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology, 62*(8), 1457–1506.
Reyes del Paso, G. A., Langewitz, W., Mulder, L. J., Van Roon, A., & Duschek, S. (2013). The utility of low frequency heart rate variability as an index of sympathetic cardiac tone: A review with emphasis on a reanalysis of previous studies. *Psychophysiology, 50*(5), 477–487.
van Rij, J., Wieling, M., Baayen, R. H., & van Rijn, H. (2022). *itsadug: Interpreting time series and autocorrelated data using GAMMs* Version 2.4.1. .
Roehr-Brackin, K. (2018). *Metalinguistic awareness and second language acquisition*. Routledge.
Rust, J., & Golombok, S. (2014). *Modern psychometrics: The science of psychological assessment*. Routledge.
Sabourin, L., Stowe, L. A., & De Haan, G. J. (2006). Transfer effects in learning a second language grammatical gender system. *Second Language Research, 22*(1), 1–29.
Salahuddin, L., Cho, J., Jeong, M. G., & Kim, D. (2007). Ultra short term analysis of heart rate variability for monitoring mental stress in mobile settings. In *Paper presented at the annual international conference of the IEEE engineering in medicine and biology society*.
Salimpoor, V. N., Benovoy, M., Longo, G., Cooperstock, J. R., & Zatorre, R. J. (2009). The rewarding aspects of music listening are related to degree of emotional arousal. *PLoS One, 4*(10), Article e7487.
Scherger, A. L. (2022). Rethinking bilingual language assessment: Considering implicit language acquisition mechanisms by means of pupillometry. *Research Methods in Applied Linguistics, 1*(2), Article 100014.
Schmidtke, J. (2018). Pupillometry in linguistic reserach: An introduction and review for second langauge researchers. *Studies in Second Language Acquisition, 40*(3), 529–549. https://doi.org/10.1017/S0272263117000195
Soares, F. C., de Oliveira, T. C. G., Tomás, A. M., Picanço-Diniz, D. L. W., Bento-Torres, J., Bento-Torres, N. V. O., et al. (2015). CANTAB object recognition and language tests to detect aging cognitive decline: An exploratory comparative study. *Clinical Interventions in Aging, 10*(37).
Soni, A., & Rawal, K. (2020). A review on physiological signals: Heart rate variability and skin conductance. In *Paper presented at the proceedings of first international conference on computing, communications, and cyber-security (IC4S 2019)*.
Steinhauer, S. R., Siegle, G. J., Condray, R., & Pless, M. (2004). Sympathetic and parasympathetic innervation of pupillary dilation during sustained processing. *International Journal of Psychophysiology, 52*(1), 77–86.
Team, R. C. (2022). *R: A language and environment for statistical computing*.
Thayer, J. F., Hansen, A. L., Saus-Rose, E., & Johnsen, B. H. (2009). Heart rate variability, prefrontal neural function, and cognitive performance: The neurovisceral integration perspective on self-regulation, adaptation, and health. *Annals of Behavioral Medicine, 37*(2), 141–153.
Thayer, J. F., & Lane, R. D. (2000). A model of neurovisceral integration in emotion regulation and dysregulation. *Journal of Affective Disorders, 61*(3), 201–216.
Thomas, B. L., Claassen, N., Becker, P., & Viljoen, M. (2019). Validity of commonly used heart rate variability markers of autonomic nervous system function. *Neuropsychobiology, 78*(1), 14–26.
Vann, R. J., Meyer, D. E., & Lorenz, F. O. (1984). Error gravity: A study of faculty opinion of ESL errors. *Tesol Quarterly, 18*(3), 427–440.
Wagenmakers, E.-J., & Brown, S. (2007). On the linear relation between the mean and the standard deviation of a response time distribution. *Psychological Review, 114*(3), 830–841. https://doi.org/10.1037/0033-295X.114.3.830
Winn, M. B., Wendt, D., Koelewijn, T., & Kuchinsky, S. E. (2018). Best practices and advice for using pupillometry to measure listening effort: An introduction for those who want to get started. *Trends in Hearing, 22*, Article 2331216518800869. https://doi.org/10.1177/2331216518800869
Wood, S. N. (2006). *Generalized additive models*. Boca Raton, Fla: Chapman & Hall.
Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B, 73*(1), 3–36.
Yoon, K. K. (1993). Challenging prototype descriptions: Perception of noun countability and indefinite vs. zero article use. *IRAL-International Review of Applied Linguistics in Language Teaching, 31*(4), 269–290.
Zeki Al Hazzouri, A., Elfassy, T., Carnethon, M. R., Lloyd-Jones, D. M., & Yaffe, K. (2017). Heart rate variability and cognitive function in middle-age adults: The coronary artery risk development in young adults. *American Journal of Hypertension, 31*(1), 27–34. https://doi.org/10.1093/ajh/hpx125
Zekveld, A. A., Heslenfeld, D. J., Johnsrude, I. S., Versfeld, N. J., & Kramer, S. E. (2014). The eye as a window to the listening brain: Neural correlates of pupil size as a measure of cognitive listening load. *NeuroImage, 101*, 76–86.