NO EVIDENCE FOR THE PREDICTIVE POWER OF THE N400 IN VOCABULARY LEARNING

Eero Uusaro
Master's Thesis
English
Department of Language and Communication Studies
University of Jyväskylä
Autumn 2022

UNIVERSITY OF JYVÄSKYLÄ

Tiedekunta – Faculty Humanistis-yhteiskuntatieteellinen tiedekunta	Laitos – Department Kieli- ja viestintätieteiden laitos
Tekijä – Author Eero Uusaro	
Työn nimi – Title No evidence for the predictive power of the N400 in vocabulary learning	
Oppiaine – Subject Englannin kieli	Työn laji – Level Pro gradu -tutkielma
Aika – Month and year Syyskuu 2022	Sivumäärä – Number of pages 58

Tiivistelmä – Abstract

Kognitiivinen neurotiede ja kasvatustiede ovat hiljattain integroituneet monitieteiseksi alaksi, jota voisi suomalaisittain kutsua opetusalan aivotutkimukseksi. Sen yhtenä kantavana ajatuksena on ollut, että aivotutkimuksesta kertynyt tieto auttaisi opetusalan työntekijöitä löytämään uusia menetelmiä saada oppijat oppimaan. Tämä vaatii mallin siitä, miten aivot tukevat kognitiivisia toimintoja.

Tässä tutkielmassa keskityttiin aikuisoppijoiden sanaston oppimisen tutkimiseen magnetoenkefalografian avulla. Aikuisoppijoille opetettiin ärsykesarjan kautta, mitkä visuaaliset objektit ja nimet kuuluvat yhteen. Tutkielmassa tarkasteltu N400 on aivoaktiivisuudesta erotettava herätevaste, jonka voimakkuuden on havaittu riippuvan siitä, kuinka paljon ärsykkeiden merkitys poikkeaa odotetusta. Mikäli N400 heijastaa opiskeltujen sanojen merkitysten tallentumista muistiin, voidaan oppimistulosten ja herätevasteen voimakkuuden ajatella kehittyvän rinnakkain.

Vastoin odotuksia ja aikaisempia tutkimuksia, voimakkuuserot N400-aikaikkunalla eivät merkittävästi ennustaneet sanaston oppimista regressiomalleissa. Empiirisen osuuden lisäksi tarkastellaan lyhyesti oppimiseen vaikuttavia tekijöitä unesta liikuntaan ja ruokavalioon. Kun tietoa koulusuoriutumisen aivoperustasta kertyy, opettajat voivat tiivistää yhteistyötään kouluterveydenhuollon ja vanhempien kanssa.

Asiasanat – Keywords	sanasto, oppiminen, aivot, aivotutkimus, MEG, magnetoenkefalografia, vocabulary, learning, brain, neuroscience, magnetoencephalography
Säilytyspaikka – Depository	Jyväskylän yliopiston julkaisuarkisto JYX
Muita tietoja – Additional information	

FIGURES

TABLES

TABLE OF CONTENTS

1 INTRODUCTION

The end of the 20th century and the beginning of the 21st century have seen the birth and growth of the interdisciplinary field of educational neuroscience. It has been defined as "the integration of education, psychology, and neuroscience into an interdisciplinary field that is devoted to helping students learn" and is thought to "communicate the language of multiple disciplines and apply methods from multiple disciplines to translate discoveries about the brain and its networks into educationally relevant outcomes" (Feiler and Stabio 2018: 23). The promise of the field has been recognized, as evidenced by research projects investigating, for example, the links of educational attainment with inhibitory control training, cardiovascular activity, and the shifting of school start times (Thomas, Ansari and Knowland 2019). The contribution of the present work is an original study of language learning, where the explicit modelling of performance on the brain response of interest, the N400, presents an improvement over previous studies. It explores central themes in educational neuroscience and suggests possible interventions. The first chapter starts with reiteration of critical discussion within the field. It then narrows the focus to concern the learning of vocabulary, and ends with an overview of the thesis.

1.1 The neuroscience of education, lost in translation?

Review of mission statements over the span of 30 years reveals that the major themes of educational neuroscience have been the application of research, interdisciplinary collaboration, and translation of the language of the multiple disciplines (Feiler and Stabio 2018). The themes give the impression that a common ground exists between its fields, which permits cross-fertilization, but only if a language barrier is overcome first. I will dub this the "translation metaphor". The metaphor of translation is interesting because it sees the fields as if they were speaking of the same problems, but were hindered by a lack of mutual understanding. This is suggested by a level view of scientific disciplines (Anderson 1972). In transitioning between the levels, social affairs are interpreted as cognitive; cognition is interpreted as brain activity, and brain

activity is interpreted as a series of electrochemical events. As will be specified below, the reality of educational practice is not simply described in the vocabulary of lower levels. Nonetheless, much interest has been directed to the study of brain functioning in these cognitively complex situations, making use of sophisticated statistical tools (e.g., Pérez et al. 2019; Bevilacqua et al. 2019).

In this work, I will consider implications from the point of view of teachers. In the Finnish context, the role of teachers is a worthwhile topic for study since, in Finland, they wield a lot of educational power (Krokfors 2017). Teachers work in conformity with the opportunities their environment offers and the limitations it poses. The curriculum and technology influence the contents, working methods and concrete exercises of instructional units. For example, vocabulary is taught through audio recordings, videos or texts provided by different publishers, the instructor or the group, and presented with the help of laptops, smartphones, projectors or, on occasion, physical books or handouts. Another key instrument of the teacher is their own voice, with which they guide the learning process and manage social matters in the group. Thus, educators utilize resources in their proximal environment, whether it is a virtual, school, home or some other environment, in various ways and for different purposes, with lessons unfolding in terms of these practical possibilities and limitations.

For interdisciplinary collaboration to thrive, it would be key that neuroscientists come to terms with such possibilities and limitations in pursuit of mutual understanding with practitioners. With this in mind, let us explore where educational neuroscience could go from here to fuel interdisciplinary work. As a thought experiment, one option would be to build a neurobiological model of some learning process that aims towards a particular outcome. The model would situate the adapting brain as a target of methods of instruction. It could describe, for example, how text reading aids the acquisition of novel words by triggering certain processes in the brain. Having identified these processes, neuroscientists would then discover how they should be manipulated in order to boost the progress of the learner. In the end, and in accordance with the translation metaphor, this discovery is translated to an educational intervention to enhance learning. The discovery, though elusive, is important, and much hope is invested in the bridge the model represents.

Even if the model were valid by virtue of being ecologically informed (Frey, Schmitt and Allen 2012), descriptive and possibly predictive success of the model does not necessarily go hand in hand with its usefulness. While the translative effort between bodies of literature in different fields promises to inform education, such an approach is not devoid of problems. Bowers (2016) has questioned the necessity of interdisciplinary cooperation with neuroscience on the basis that both educational interventions and assessment happen on a behavioral level, and theorizing about behavior belongs to the field of educational psychology. This is different from recent

criticism that aims to counter neurocentric reductionism (Krakauer et al. 2017), and not attempt to undermine the crucial role of the brain in virtually all human experience, but to point out that innovations from neuroscience are limited by practical matters. The fact that brain activity changes during classroom learning seems little consequential to the teacher. For this reason, the disciplines would not complement each other.

There is, however, a corollary to the position, in which the neuroscience of education may broaden the perspective on interventions by suggesting some that might be unintuitive in a purely psychological approach. In addition to a model of learning, this requires interpreting in neuroscientific terms the effect of interventions outside psychology, so to discover links between them and educational attainment. Though out of question, in the extreme example, it could mean the administration of nootropics, colloquially known as smart drugs, to healthy with the intention to boost memory. However, possible interventions could include more ordinary ways of cognitive improvement relating to nutrition, physical activity or sleep, though Bowers (2016) equally questions their novelty value. I return to these issues in the Discussion, where I consider the role such topics could occupy in this scheme, and whether there could be room for cooperation between school healthcare and teachers. In summary, this conception sees neuroscience as the bridging field between multiple disciplines, while there is less need for a bridge to neuroscience in specific.

1.2 Not a property of the words: meaning in language education

The present work concerns language education, so it makes sense to briefly consider the benefits that command of a language yields. Before beginning, though, it must be noted that varying practices of language use do not always agree with common-sense definitions of a language. Definition in terms of mutual intelligibility or geographic borders is problematic, as there are a number of exceptions (Holmes 2013: 137–138). Thus, it is probably for the best to loosely talk about semiotic resources used for different purposes (van Lier 2004: 96–97), which can also be referred to with working terms like English. While it is possible to use language non-communicatively, for example by externalizing thoughts to plan actions, language skills and knowledge crucially allow one to voice views in a community, draw from the pool of knowledge belonging to that community, and manage social affairs within it. For example, scientific knowledge is chiefly distributed in English, and the central place English occupies in the technology sector, economy and popular culture is undisputed (Crystal 2004). Therefore, language education can be very empowering.

Semiotic resources refer to possibilities for sign use. Though the centrality of what linguistic signs stand for seems obvious, let us revisit the views of other authors on sign systems, lest the discussion become self-contained. The work of Hockett (1958: 569–586), which I do not delve very deep into, should suffice to this end. I use the criticism directed towards it also to justify the explicitly cognitive view endorsed here. Over sixty years ago, Hockett proposed that human language exhibits a group of design features, notably arbitrariness and displacement: an account of lasting influence in linguistics (e.g., Yule 2017: 13–17). Displacement means that linguistic expressions can refer to outside of their reception context. Moreover, they need not display a natural connection to or resemble what they stand for, that is, their meaning can be arbitrary. Without displacement and arbitrariness, our communication would be very limited and more cumbersome. Despite their influence, design features have been criticized for their assumed role as properties of linguistic signals, which neglects the cognitive and social abilities of the language user that make them possible (Wacewicz and Żywiczyński 2015).

This neglect is true for word meanings as well. Like language, what a word is resists common-sense definitions, such as those with basis in word form (Carter 2012: 20–23). To the dictionary editor, homonymous words are distinguished by their meanings, yet due to context-dependent polysemy it is difficult to identify essential meanings. Thus, Hanks (2000) proposes that words have meaning potential that is realized situationally. From this point of view, meanings are events rather than entities. I will mention in the Theoretical Background, how, in fact, such probabilistic accounts of meaning have proved very effective (Rabovsky, Hansen and McClelland 2018) in predicting brain activity that relates to semantic congruity (Kutas and Federmeier 2011). The event-oriented view invites commentary similar to that of Wacewicz and Żywiczyński (2015) on Hockett's design features. Cornejo (2004) offers some in his discussion of psycholinguistics where, on one hand, word meanings are held to be independent of the subject; on the other, constructionist views in psychology stress that they result from active interpretation on the part of the individual.

The purpose of the discussion so far has been to draw attention to the cognitive and social abilities of the language user. Insofar as the brain is concerned, it is arguably best understood in the context of the behavioral problems it must solve (Krakauer et al. 2017). Here, the behavioral, or computational (Marr 1982: 8–38), goal is to maximize linguistic expressivity. Taken as feats accomplished by cognition, displacement and arbitrariness could be seen as algorithmic solutions to this end, to borrow Marr's (1982: 8–38) terminology. The position assumed in the present work is that their implementation occurs in integration with brain systems that serve varieties of declarative memory. Although I focus on the role of memory, memory must be considered in conjunction with social factors. As already argued by Wittgenstein (1953), linguistic

expressions need to have meaning to other people in order to be considered language. Indeed, the tendency to regard communication signals as honest, even across kin, has been presented as characteristic of human language (Wacewicz and Żywiczyński 2018).

1.3 Interim summary and structure of the thesis

Statement of objectives and general theoretical dispositions is in order. First, focus is reserved for the cognitive abilities of the language user, instead of the expressions they use (Wacewicz and Żywiczyński 2015). Second, this is not to neglect the idea that meaning is a theoretical concept used to explain intersubjective agreement on the use of linguistic expressions, as well as the interpretive recreation of their contents in the individual mind (Cornejo 2004). Third, although the experimental paradigm, described in the Methods, could be seen to concern a fictional language, the study is not held to be specific to any particular language, insofar as no principled way to mark boundaries between them exists (Dufva et al. 2011). Rather, the work regards language as semiotic resources (van Lier 2004: 96–97), which seemingly display a number of characteristics that contribute to expressivity. Fourth, it pursues a model of learning to discover how the brain supports the language user in achieving this goal. Fifth, the significance of the model for educational neuroscience lies in designing non-psychological interventions (Bowers 2016). A pragmatic point of view is advocated in their application.

This paragraph marks the end of Chapter 1. The Introduction should now have identified the fields of study where the work belongs to, described some running themes, and outlined relevant research areas. Next, I examine word meanings in greater detail. Chapter 2 starts with a discussion of the relationship between behavior and the brain. It then describes how the brain supports word comprehension. From there on, the work focuses on the role of a singular brain response, the N400, in the processing of meaning, interpreted in a framework titled Controlled Semantic Cognition (Lambon Ralph et al. 2017). The chapter ends with a comprehensive review of the N400 in studies of second language learning. Chapter 3 describes an original N400 study, which utilizes magnetoencephalography and regression modelling to predict learning outcomes from the response. Chapter 4 presents the results of the study. Chapter 5 discusses their implications and, in a thematic fashion, outlines neuroscientifically inspired interventions. Chapter 6 presents concluding thoughts, where I reconsider the main tenets of educational neuroscience, and the role of knowledge accumulating from the field.

2 THEORETICAL BACKGROUND

This chapter describes and discusses a measure of how the brain processes meaning, known as the N400 (Kutas and Federmeier 2011). The N400 is an offspring of brain research using electroencephalography (EEG) and magnetoencephalography (MEG). These brain imaging methods, whose operating principles are described in the Methods, track brain activity with high temporal resolution. One way of using M/EEG has been to investigate brain responses that are time-locked to some event of interest. For example, the presentation of a visual stimulus to the subject can be time-stamped on the M/EEG data that is collected at the same time. Comparison of brain responses when the visual or some other stimulus is experimentally manipulated yields effects in different time periods (Luck 2014: 71–118). These effects and the related deflections in the M/EEG time series are termed event-related potentials (ERPs) in EEG research and event-related fields (ERFs) in MEG research. The common term event-related responses, under which the N400 also belongs, is used to refer to both.

The first section brings up and discusses important conceptual issues. These relate to limitations of brain imaging modalities and to the primacy of behavior in measuring psychological constructs, understanding the role of the brain, and as the target of education. Word meanings have to be learned on the basis of experience and are thus memories. The second section describes the declarative memory system, while the relationship of the N400 to semantic processing is considered in the third section. Discussion of its sensitivity to experimental manipulations shows it to be more than just a neurolinguistic phenomenon. An account of the N400 is given that draws upon simulation and localization work, and is situated within the larger framework of semantic memory. The final section closes the chapter with a comprehensive review of the N400 in second language (L2) learning and proficiency studies. This review highlights the fact that performance in learning paradigms has not been modelled on the N400, although this would seem important for educational purposes.

2.1 Minding behavior and the brain

The first chapter addressed issues in educational neuroscience as if the brain were open to inspection and without consideration for what constructs like learning and memory actually are. This section devotes a brief discussion to these topics, starting from brain imaging modalities and their differing fields of view. The abundance of imaging modalities (Purves et al. 2012: 18–20) suggests that they are complementary, each making up for the shortcomings of the others. Otherwise, there would be little point in using many imaging tools to probe the brain. For instance, although M/EEG provides the best temporal resolution out of all non-invasive modalities, not all brain activity is readily shown. In the case of MEG, as the signal-to-noise ratio worsens as a function of source depth, data acquisition and processing must be accordingly optimized to distinguish activity originating in deep sources such as the hippocampus (Ruzich et al. 2019), which is routinely accomplished in functional magnetic resonance imaging (fMRI). The same point applies to data analysis. In contrast to evoked responses, brain oscillations might lead to different conclusions about how the brain functions in a given situation. Consequently, brain activity and measurements of it must be separated conceptually, which helps reconcile evidence from different imaging modalities.

The notion of a brain functioning in a given situation demands elaboration. It is tempting to treat these functions as equal to cognitive processes like perception, recognition or comprehension. However, this is not the case; instead, such processes are theoretical concepts that are used to explain performance in behavioral paradigms, which have been designed to isolate them. For example, in investigating memory, the participant might perform a test where a set of items must be memorized. Following Rudy (2014: 154–157), any neurophysiological processes that might be measured in this paradigm are secondary to performance in determining which of them count as memory processes. Thus, if no change in performance has taken place, but brain activity has changed somehow, no implicit learning has occurred. This parallels the contention that education should not lose sight of behavior as its principal target in favor of the brain (Bowers 2016). Nevertheless, the role of cognitive processes as theoretical concepts does not mean that they are devoid of physical reality. All behavior is a physical feat that requires a functioning brain, and the brain can be seen as supporting the performance that is then used to measure the latent constructs.

As stated previously, the promise of educational neuroscience is that this supporting role may be understood well enough to widen the perspective on learning, possibly by transforming insights from the health sciences into improved educational attainment. However, what is meant by good enough understanding has not been clarified. One approach is to consider the issue in terms of biomarkers and surrogate

outcomes. A biomarker is any objective measure of the biological structures, substances or processes of the body or its products, while a surrogate outcome is obtained as a substitute for some true outcome, often with the purpose of predicting it before it is observed (Aronson 2005; Strimbu and Tavel 2010; Baker and Kramer 2013). A biomarker can function as a useful surrogate without displaying a solid biological link with the true outcome (Strimbu and Tavel 2010; Baker and Kramer 2013). In fact, its utility depends on its statistical connection to the outcome and whether it is practical to obtain (Baker and Kramer 2020). For example, knowing that selective histone deacetylase inhibition improves learning and memory in healthy rodents by promotion of DNA transcription (Gräff and Tsai 2013) may be useful, whether or not learning and memory amount to it. I will return to this idea in the review of studies on L2 learning and the N400.

2.2 Memory and word meanings

Word meanings are not given but must be learnt over time, which calls for a discussion of the memory system that supports this. In general, memory is a theoretical term used to explain the fact that behavior is influenced by experience (Rudy 2014: 2). Cohen and Squire (1980) originally proposed a division between declarative and procedural memory to accommodate the finding that amnesic patients could acquire a mirror-reading skill while performing poorly in recognition memory tests. In this division, a defining property of the former is the conscious recollection of information (Squire 2004), therefore encompassing two types of memory distinguished earlier, episodic and semantic memory (Tulving 1972). Episodic memory is contextual, which means that activation of episodic memory traces leads to mnemonic reinstatement of the situation where the memory trace was formed. Thus, seeing a yellow umbrella would bring back memories of the woman who was holding it while walking her dog on a rainy day. In contrast, semantic memory does not include reference to the context in which the memory traces were formed. Therefore, the knowledge that Helsinki is the capital of Finland might not contain information about where such a fact was learned.

In both episodic and semantic memory, pieces of information are bound together such that cued retrieval can occur. The brain accomplishes this through convergent inputs of neurons. For instance, episodic memory has been argued to work by conjoining inputs in medial temporal-hippocampal structures (Rudy 2014: 309–311). To build upon the previous example, broad neocortical activity constitutes the experience of a woman walking her dog on a rainy day. Inputs from these neurons converge on downstream neurons to form a compressed representation of the experience. A perceptual cue such as the yellow umbrella can activate these downstream neurons, and

if memory traces have been formed in them, the original neocortical activity can be reinstated. In contrast to episodic memory, semantic memory shows little hippocampal involvement, and is instead supported by neocortical convergence zones (Binder and Desai 2011; Renoult et al. 2019). In the complementary learning systems (CLS) theory, the formation of semantic memory traces in such structures results from the extraction of structured knowledge embedded in experiences and episodic memories (McClelland, McNaughton and O'Reilly 1995; Kumaran, Hassabis and McClelland 2016), as illustrated in an analogous fashion in Figure 1.

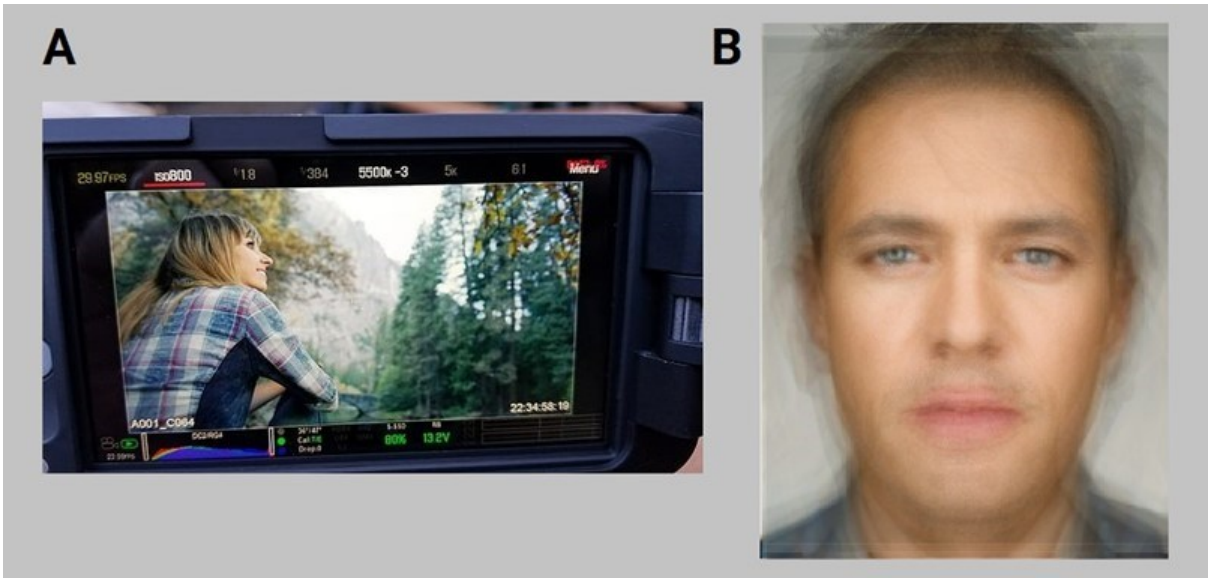

FIGURE 1 Whereas episodic memory can be likened to a video recorder (A) that faithfully captures little details, semantic memory is more akin to a composite portrait (B), working to extract structured knowledge in the environment. The fast, auto-associative binding and storage of the information that makes up individual experiences is thought to be supported by the hippocampus, while structured knowledge is argued to be gradually stored in the neocortex (McClelland et al. 1995). The formation of the latter is influenced by hippocampus-induced replay of stored experiences, which may be biased towards those that involve novelty, high informational value, or a reward (Kumaran et al. 2016), and which occurs during periods of rest, including but not limited to sleep (Klinzing, Niethard and Born 2019). Furthermore, the learning of structured knowledge is faster when it is consistent with existing prior knowledge (Kumaran et al. 2016). A adapted from Unsplash (https://images.unsplash.com/photo-1509555163580-ff2bd082c8e0) under their license (https://unsplash.com/license). B adapted from https://flic.kr/p/cNaJGN under the CC BY-SA 2.0 license.

2.3 Meaning in language in the light of the N400

The N400 was first discovered by Kutas and Hillyard (1980) in an EEG experiment where subjects had to silently read sentences in order to answer questions about them later. These sentences were presented one word at a time and differed in their endings. While most of the sentences ended semantically appropriately (e.g., *She put a stamp on the letter*), some endings were inappropriate (e.g., *She put a stamp on the window*) and some outright nonsensical (e.g., *She put a stamp on the electricity*). When the difference between ERP waveforms to the endings of the semantically incongruent and congruent sentences was measured, a prominent negativity would show around 400 milliseconds post-stimulus for the incongruent sentences. This effect, which is canonically quantified as the difference between mean amplitudes from 300 to 500 milliseconds post-stimulus, was subsequently called the N400. An N400 effect with the same waveform and scalp topography is produced by contrasting ERPs to words preceded by semantically related or unrelated word primes, for example by contrasting those elicited by *boat* when it is primed by *ship* or *serpent* (Holcomb 1993; Kutas 1993).

At least partial similarity in how the brain processes meaning in language and outside of it is supported by N400 studies in non-linguistic domains. For instance, larger N400 amplitudes are produced by semantically incongruent endings to videos (Sitnikova et al. 2008), visual objects that are semantically incongruent with the scenes where they appear (Ganis and Kutas 2003), and semantically incongruent environmental sounds (Schirmer et al. 2011). A domain-crossing paradigm has been used in a number of mostly clinical studies, where visual objects and names are paired, and in some of which participants have to verify the correctness of these pairings. These studies have shown larger N400 amplitudes to be produced for words not preceded by their referent in healthy (control) populations of various ages (Stelmack and Miles 1990; Ford et al. 2001; Mathalon, Faustman and Ford 2002; Mathalon, Roach and Ford 2010; Kuipers, Jones and Thierry 2018). Not only do these studies corroborate the idea of domain-general processing of meaning, but their paradigm is also adaptable to learning studies probing the possibility that the development of the N400 follows the emerging connections between objects and their names, or vice versa.

Remarkably, all of the main findings from N400 studies have been reproduced by simulating it as a computational process in the sentence gestalt (SG) model (Rabovsky et al. 2018). The SG is a representation of compositional meaning, consisting of semantic features in probabilistic activation states. Word inputs to the SG change not just the probabilistic activation of their own semantic features, but also those of statistically co-occurring, subsequent words. For example, the inputs *John walks* will activate features that describe their referents, such as male and movement,

and also those relating to possible continuations of the sentence such as pet and canine for *his dog* or long duration for *slowly*. Either of the continuations will more strongly activate some features while attenuating the activation of others in a process termed the semantic update, the magnitude of which has been found to closely follow the N400 in simulations of prior studies. This agrees with an earlier conclusion that the N400 is a process and time period in which the brain tries to reconcile the outcomes of perceptual processing with the state of semantic memory (Kutas and Federmeier 2011).

2.3.1 Anatomical interpretation of the SG model

The reasons why a construct like the SG could be supported by the anterior temporal lobe (ATL) are twofold. First, in a clinical condition termed semantic dementia (SD), progressive brain degeneration that begins in the ATL results in a loss of conceptual knowledge (Lambon Ralph et al. 2017). Typically, SD patients tend to overgeneralize frequent and typical concepts, while untypical concepts are undergeneralized or not used at all (Patterson and Lambon Ralph 2016). For example, when asked to name a zebra shown in a picture, SD patients might call it a horse, whereas a platypus might be unnamable to them. Second, the ATL displays a combinatory effect such that nouns produce stronger activation when preceded by modifiers, in contrast to meaningless consonant strings (or listed items) (Pylkkänen 2019, 2020). This effect is stronger for low specificity nouns like *boat* in contrast to high specificity nouns like *canoe*, when modified by words such as *blue* or *long* (Westerlund and Pylkkänen 2014), and for low specificity nouns that are preceded by highly specific nouns in contrast to low specificity nouns to form compounds (Zhang and Pylkkänen 2015). The results from these two lines of research fit well with the idea that the ATL is a brain structure supporting the SG.

The combinatory findings have been explained by an account in which ATL activation follows the relative contribution of the first item to the semantic feature set of the composition (Pylkkänen 2019, 2020). Assuming that consonant strings produce noise in the ATL, the account also accommodates the weaker evidence that specific nouns activate it more strongly than the general, when both follow consonant strings (Westerlund and Pylkkänen 2014; Zhang and Pylkkänen 2015). A compatible view is that general concepts activate the ATL more broadly, as they include realizations that fall under them. For example, the realizations of boat include canoe but also yacht and ferry. A given modifier (e.g., *long*) produces intersection with this activation and requires inhibition of irrelevant activation, which could determine ATL signal amplitudes. Though less parsimonious an account, it is favorable because the SG likewise involves cohort-like activation of possible meanings by single inputs. This account also agrees with the finding that higher ATL concentrations of the main inhibitory

neurotransmitter γ-aminobutyric acid (GABA) are associated with better performance in a semantic association task (Jung et al. 2017).

In the semantic update, new inputs change probabilistic activation within the SG. There are two reasons why this process might be carried out in part by the posterior middle temporal gyrus (pMTG) and the inferior frontal gyrus (IFG). First, the clinical condition associated with their lesions is marked by impaired semantic control rather than storage. For example, upon seeing an ambiguous word such as *trainers*, denoting a type of shoes or people involved in the act of training, SD patients are more likely to demonstrate a loss of any such conceptual connections, while IFG or pMTG dysfunction will impair the ability to suppress one of the meanings in favor of the other (Nozari and Thompson-Schill 2016). These symptoms suggest that pMTG and IFG work together to shape activation patterns in the ATL. Second, a review shows them to be the principal candidates for the source of the N400 (Lau, Phillips and Poeppel 2008). Though the authors argue mainly for the involvement of the pMTG in this regard, they note that N400-like effects have also been measured from anterior temporal regions. Altogether, these results agree with the idea that the N400 represents a change in the probabilistic activation of semantic features, and suggest it to be related to controlled retrieval. The framework where the ATL is the conceptual storage, pMTG and IFG being control regions has been titled Controlled Semantic Cognition (CSC, Lambon Ralph et al. 2017), depicted in Figure 2.

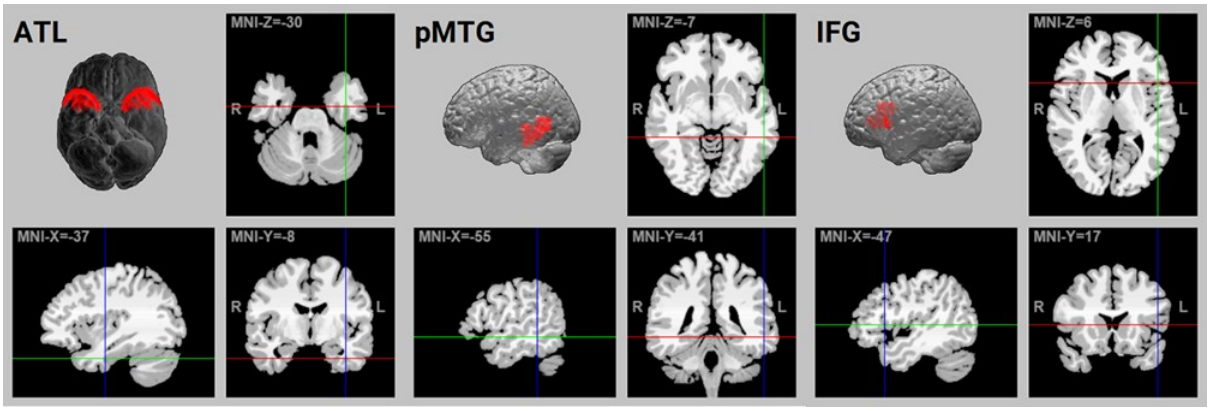

FIGURE 2 CSC regions. It is worth mentioning that SD symptoms correlate best not with the very end of the temporal lobe, but the region slightly posterior to it (Lambon Ralph and Patterson 2016). Red highlights mark the locations of the three regions on the basis of Brodmann areas (BAs). The ATL includes part of BA 38, while the pMTG corresponds to the posterior part of BA 21. Only the posterior section of BA 21 is highlighted. In turn, BA 45 represents the IFG. The brain images with red highlights were adapted from Wikimedia Commons under the CC BY-SA 3.0 license. The black panels with Montreal Neurological Institute coordinates of the regions were produced with the help of the MNI2Tal tool (https://bioimagesuiteweb.github.io/webapp/mni2tal.html), part of the Yale BioImage Suite.

2.3.2 CSC regions jointly produce the N400

The ATL, pMTG and the IFG work together to process the meaning of objects in different contexts. For example, semantically incongruent sentence endings produce both greater N400 amplitudes in the pMTG and greater suppression of beta oscillations in the IFG, although within the incongruent condition, less beta suppression predicts greater N400 amplitudes (Wang et al. 2012). The anatomical connectivity of the three regions is supported by two white matter pathways connecting the temporal lobe with the frontal: an indirect pathway formed by inferior longitudinal and uncinate fasciculi, and a direct pathway formed by the inferior fronto-occipital fasciculus (IFOF) (Duffau et al. 2014; Duffau 2015). Notably, direct electrical stimulation (DES) and diffusion MRI studies provide evidence for the involvement of each white matter pathway in semantic processing which, however, is strongest for the IFOF (Cocquyt et al. 2020). In these studies, DES of the IFOF consistently produces semantic paraphasias, and its diffusion metrics correlate with performance in comprehension and semantic production tasks.

The temporal dynamics of semantic processing in the three regions is illustrated by a study into their oscillatory responses by Teige et al. (2018). Although the authors were primarily interested in contrasting associative strength in word pairs, their supplementary analysis focused on word relatedness. Contrasting word relatedness, three successive effects in the lower gamma range emerge within 200–350 milliseconds post-stimulus, possibly in an effort to interpret mismatching ATL activation patterns: first in the IFG, then in the ATL, and finally in the IFG to the other direction. As the group used visual stimuli, the first effect is likely mediated by the IFOF, which provides a direct pathway from the visual regions to the frontal. In turn, the ATL effect corresponds in time with the combinatory ATL effects referred to above (Westerlund and Pylkkänen 2014; Zhang and Pylkkänen 2015). The three effects are followed by an effect in the beta range in the pMTG within 350–400 milliseconds post-stimulus, which would coincide with the N400 peak. Finally, the ATL shows an effect in the alpha range in a time period beginning at 400 milliseconds post-stimulus, during which chronometric transcranial magnetic stimulation applied to the ATL of healthy subjects interferes with performance in semantic relatedness judgment (Jackson, Lambon Ralph and Pobric 2015).

In sum, this view of the N400 predicts that, aside from inputs to the SG, the response is affected by integrity of the tripartite CSC system and plasticity affecting ATL activation patterns. As part of the neocortical memory system, such plasticity in the ATL should occur according to the principles of interleaved learning, as in the CLS theory (McClelland et al. 1995; Kumaran et al. 2016). Furthermore, the above results put to question the necessity of localizing the N400 to a single region. Damage to either the IFG (Friederici, von Cramon and Kotz 1999) or the ATL (Kotz, Opitz and Friederici

2007) diminish the N400 effect, and while an abnormal N400 is associated with ATL hypometabolism in mild SD (Grieder et al. 2013), evidence from source localization and other methods points to the pMTG, though not unanimously (Lau et al. 2008). Thus, it may be more reasonable to speak of the network underlying the N400, which would be in line with a network-oriented view of language functions (Duffau et al. 2014), or cerebral functions in general (Duffau 2015).

2.4 The N400 in L2 semantic learning and proficiency studies

This section describes and discusses longitudinal, cross-sectional and mixed longitudinal studies that explore the relationship of the N400 response with semantic performance in L2. These studies have almost unanimously shown that differences in performance coincide with amplitude differences in the N400 (Soskey, Holcomb and Midgley 2016; Ojima et al. 2011; Pu, Holcomb and Midgley 2016; Yum et al. 2014; Phillips et al. 2004; Elgort et al. 2015), or its latency (Ojima, Nakata and Kakigi 2005). In only one study were performance differences not found, even when N400 amplitudes differed (McLaughlin, Osterhout and Kim 2004). It is convenient to think of the studies as falling into three groups, which are termed here learner follow-ups, studies with training protocols, and proficiency group comparisons. To illustrate their defining qualities and differences, some exemplifying studies are singled out for description in the following. A couple first language (L1) studies are included to give examples of alternative ways in which the relationship between semantic learning and proficiency and the N400 has been investigated. Several critical issues are then raised and discussed in order to provide context for the present study.

2.4.1 Overview of studies

Participants in the first type of longitudinal studies, learner follow-ups (McLaughlin et al. 2004; Soskey et al. 2016; Ojima et al. 2011), receive instruction in the target language outside the laboratory and have their evoked responses measured over the course of the process. The experimental paradigms used in their elicitation typically make use of semantic priming and congruent or incongruent sentence contexts in the language to be learned. The performance metrics derive either from laboratory-based tasks or external sources, such as the institution providing the education. Behavioral and neurophysiological analyses then proceed side by side. The learning is supposed to occur entirely outside the laboratory, which distinguishes learner follow-ups from the second type of longitudinal studies, studies with training protocols. In addition, the follow-up periods exceed those of the latter in duration, ranging from months to years. However, both take advantage of instruction with purpose of observing how

the development of the N400 and changes in performance might co-occur over time, longitudinal contrasts forming the basis for analysis.

An exemplifying study is provided by Ojima et al. (2011), who followed primary schoolers of low, medium and high proficiencies over the course of three years, thus adding a cross-sectional aspect. Each year, their progress in the target language (English) was gauged by a listening comprehension test, and their ERPs were measured in a paradigm where a set of spoken words were primed by semantically congruent or incongruent pictures. Different developmental stages could be discerned over time and across proficiency groups, in which a broadly distributed negativity to incongruent words appears and shifts to an earlier time-window to become the N400 response. In the final stage, a late positivity appears in posterior regions, resembling the response when the linguistic stimuli were presented in the mother tongue (Japanese). The study suggests it to be instructive to look at epoch-wide response, instead of fixating on a single component like the N400. While the ERPs to semantically primed spoken words developed, the group did not control for prior knowledge of the words, which presents a shortcoming, possibly confounding the results.

Studies with training protocols (Pu et al. 2016; Yum et al. 2014) differ from learner follow-ups in that the subjects undergo laboratory-based training. In one study, four hours of association pairing of L2 (Spanish) and L1 (English) words resulted in improved L2-to-L1 translation performance (Pu et al. 2016). This was accompanied by a more negative N400 to non-translations in comparison to translations in the L1 when these were primed by the L2 words. During the training, L2 words were read out aloud to the subjects, who orally repeated them. In another study, participants had to learn L2 (Chinese) words over the course of ten days (Yum et al. 2014). They received daily training on word perception and translation tasks in a controlled environment, EEG measurements taking place every third day. Performance in L2-to-L1 translation was used as behavioral measure. The learners were then divided into two proficiency groups using a median split for performance in the task. The resulting high proficiency group outperformed the low proficiency group in every behavioral measurement and showed more negative N400 response to unpaired L2 words beginning in the second session.

Proficiency group comparisons (Phillips et al. 2004; Ojima et al. 2005; Elgort et al. 2015) contrast the evoked responses of groups distinguished by some metric of linguistic proficiency. For instance, the N400 was found to occur earlier for L1 speakers compared to Japanese L2 speakers, and earlier for a high proficiency group compared to a low proficiency group, when elicited in an English version of the sentence context paradigm (Ojima et al. 2005). The groups were separated by their scores in the Test of English for International Communication. The result is reminiscent of the follow-up of primary schoolers, which found a similar latency shift in the negative potentials as

performance improved (Ojima et al. 2011). Moreover, these two studies demonstrate how, in the better performing Japanese learners, the N400 responses for the L2 begin to resemble not only their own L1 responses, but also those of native speakers. This is in line with a review of hemodynamic and electrophysiological studies showing that processing differences in the L1 and L2 can largely be accounted for by differences in proficiency and age of acquisition (Indefrey and Davidson 2009).

Comparable L1 studies showcase design choices that have not been implemented in L2 research. Batterink and Neville (2011) studied implicit word learning in L1 and had their subjects judge the semantic relatedness of words in a task where the ones to be learned primed known words. The novel words were learned from sentences, instead of translations or definitions. The participants showed an N400 effect to words primed by recognized, but not unrecognized novel words, thus demonstrating a memory effect. The presumed relationship between the N400 and performance in semantic tasks has also been presented more explicitly in L1 studies. For one, Coch and Benoit (2015) regressed performance in a measure of receptive vocabulary, the Peabody Picture Vocabulary Test (PPVT), on N400 amplitudes. In PPVT, the test administrator says a word and presents four pictures to the participant, who then selects the picture they think matches the word in meaning. More negative N400 amplitudes to unpaired L1 words in children were found to be associated with higher test scores, although they explained little of the score variance.

2.4.2 Room for improvement

In this sub-section, I outline shortcomings in the reviewed studies for this and future work to improve upon. First, the learner follow-up studies (McLaughlin et al. 2004; Soskey et al. 2016; Ojima et al. 2011) unfortunately did not document the methods of instruction that ultimately both produced the N400 and improved performance in all but one of the studies (i.e., McLaughlin et al. 2004). The educational institutions come off as black boxes, which does not help in designing interventions to speed up learning. Studies with laboratory-based training protocols, however, were sufficiently specific about their instruction. Starting off with Soskey et al. (2016), the methods included exposure to L2 words, L1-to-L2 association pairing and two-choice L1-to-L2 translation with feedback. Pu et al. (2016) made use of exposure to L2 words, their oral repetition, bidirectional association pairing and two-choice translation of L1 and L2 words with feedback, while Yum et al. (2014) had their subjects detect word repetitions, pair L1 words with L2 words, judge the correctness of translations, and produce the L1 translations of L2 words. Comparisons of, say, the effectiveness between the methods would be desired.

The idea could be extended to contrasting the effectiveness of wholly different types of instruction. A systematic review of vocabulary instruction methods to

increase reading comprehension shows mere passive exposure to words and their meanings to be less effective than tasks that prompt one to make use of the meanings by, for instance, comparing, applying or analyzing them (Wright and Cervetti 2016). At the opposite end is incidental word learning, where meanings are learned implicitly from context, and which in N400 studies has been explored in parallel L1 learning literature. For example, Batterink and Neville (2011) had their subjects read through L1 text where novel words either consistently or inconsistently replaced other words. In L2 acquisition studies, the paradigm remains largely unexplored. Future studies probing the association of changes in the N400 with L2 semantic learning could therefore compare the effects of passive exposure to and active processing of the stimuli. Nevertheless, it is clear that studies should start from better specification of the instructional methods, especially if the instruction is offered by an educational institution.

Second, in part, the studies reviewed in this section are precursors to educational neuroscience, following more closely the older tradition of N400 studies discussed earlier. Possibly owing to this, they were not sufficiently specific about the relationship between the N400 and performance for educational applications to be considered. As stated in the Introduction, a way to apply the knowledge is to devise interventions that target the brain so as to improve educational attainment: what is required is a model of learning that describes the influence of the brain on performance, not vice versa. Typically, the studies used both performance and brain activity as outcome variables, which permits only limited inferences. When they did not, performance formed the basis for comparison of brain activity. Coch and Benoit (2015) provide an exception to this, though in L1 literature. All in all, not much attention is paid to the relationship between performance and brain activity at the conceptual level. While this criticism is generally valid for neuroscientific studies (Krakauer et al. 2017), the problem is all the more important to applied research.

One option is to use the N400 as a biomarker and surrogate outcome for learning. It could be construed as an index of microstructural changes in the brain that support learning (Sweatt 2016), but without a strict need for clarifying the relationship biologically (Strimbu and Tavel 2010; Baker and Kramer 2013). This seems questionable because measuring the primary outcome directly is cost-effective and straightforward in comparison to collecting brain data, unlike in, say, cancer follow-up studies. Finally, changes in the brain may precede those in behavior, in which case evidence of behavioral change over time is required. It does not follow from mere null effects in task performance that learning has occurred if the brain has changed. Nevertheless, such conclusions have been drawn from the observed neurophysiological and null behavioral effects in Laughlin et al. (2004): "significant cognitive learning had occurred before behavioral data showed any difference" (Soskey et al. 2016: 45). This view has

been forcefully rejected by Bowers (2016): without appropriate behavioral change, speaking of anything cognitive is not warranted.

3 METHODS

Investigation of the neurophysiological processes underlying the learning of word meanings involves a number of methodological questions. These relate not only to how brain activity and learning are measured, but also to how the dependence of one on the other is determined. This chapter outlines a suitable methodology by beginning with a brief discussion of the place of MEG in learning research. The aims of the present study are then clarified in more detail. Following that, the next section describes the profile of the subjects and their preparation together with relevant ethical considerations. The structure of the experiment is then described, followed by a description of the procedures to reduce noise in the data. Finally, the last two sections discuss and describe the statistical analysis strategy. That a broad discussion of brain imaging methods precedes all this might seem overly general, but is necessary as experimental design and imaging methods are deeply interwoven.

3.1 Magnetoencephalography in learning research

It was argued in the Introduction that the strength of adopting a neurobiological level of explanation lies in making it easier to understand the contribution of psychologically unintuitive factors to behavior. This means understanding the theoretical concepts of perception, comprehension and learning empirically, in neurophysiological terms. There are a number of brain imaging methods suitable for this purpose, which include but are not limited to fMRI, EEG and MEG. They differ in the way brain activity is measured; for example, there may be differences in whether the measurements concern cerebral metabolism, or more directly the communication between neurons (Purves et al. 2012: 18–20). Temporal resolution is a particularly important performance metric in the context of learning research. In many learning paradigms, such as the ones described at the end of the Theoretical Background, the

neurophysiological changes involved in learning occur between a stimulus and the response to it. MEG emerges as the imaging method of choice when it is important to describe brain activity in such a time window, for reasons that follow.

In neuronal communication, presynaptic activity results in electric current flow across the postsynaptic membrane, which, by changing the electric potential of the membrane, either brings the postsynaptic neuron closer to its firing threshold, or distances it from it. The summation of multiple excitatory or inhibitory postsynaptic potentials co-occurs with a larger current flow within the postsynaptic neurons (Purves et al. 2012: 102–105). When a sufficiently large number of adjacent postsynaptic currents are present simultaneously, their summation may be detected by means of EEG or MEG. Whereas EEG is predicated on the spread of electric current through the surrounding tissues, MEG makes use of the fact that the current generates measurable magnetic fields. Because the head is almost transparent magnetically, the relative strength of MEG in this regard is that the tissues produce less signal smearing. Furthermore, in MEG measurements, it is primarily the pyramidal cells in cortical sulci, that is, currents that are oriented tangentially relative to the head surface that contribute to the measured signal, due to the orientation and anatomy of these cells and the orientation of magnetic fields relative to the currents that produce them (Hari and Puce 2017: 25–37). Unlike in EEG, radial currents are rather invisible.

The contribution of the currents to the measured signal is virtually instantaneous, which gives MEG its very high, sub-millisecond temporal resolution. In this regard it outperforms fMRI, in which measurements correspond to changes in cerebral blood oxygenation that happen due to changes in brain functioning. The high temporal resolution complements behavioral measures used in psychology such as reaction times, which are end state measures in that they reflect all the processes leading up to them. MEG is able to track ongoing brain activity before the response, and with appropriate experimental manipulations will provide a more fine-grained picture of the processing steps involved. Due to its origin, the MEG signal is also a more direct measure of the functioning of neurons. On the other hand, it is problematic to make inferences of the source of the measured MEG signal, since there are an infinite number of sources that could produce it. This is known as the inverse problem in MEG research. Unlike in fMRI, the anatomy of the head is not measured in MEG. Rather, the two main sensor types, gradiometers and magnetometers, are sensitive to local and remote sources, respectively. Solutions to the inverse problem therefore involve making assumptions about the source space in order to narrow down the number of possible sources. Nevertheless, when the purpose is not to specify the sources that generate the signal, MEG emerges as the imaging method of choice to study learning.

3.2 The present study

The present study is an MEG study that aims to contribute to research on the N400 by testing the power of the semantic congruence effect to predict performance in semantic decision. As stated in the preceding chapter, this approach reverses the typical analysis procedure of learning and proficiency studies where the N400 is considered to be the dependent variable (McLaughlin, Osterhout and Kim 2004; Batterink and Neville 2011; Yum et al. 2014; Soskey et al. 2016). Furthermore, the aim is not only to characterize the relationship of the N400 effect to performance, but also to investigate the parallel development of performance and neurophysiological effects in time. Thus, the presence of a semantic congruence effect was first assessed across time and the size of the effect was then used to predict performance in semantic decision. More generally, this study attempts to construct a neurophysiologically motivated model of performance in semantic decision that can be used to inform and inspire educational practices in the larger context of educational neuroscience.

Such an approach calls for a combination of analysis methods. Following data acquisition and noise reduction, exploratory significance testing of effects in the N400 time window was combined with generalized least squares (GLS) estimation to determine *whether the size of the effect predicts performance in semantic decision*. If so, the second question is *how well this performance is predicted*. In its analysis strategy and use of MEG, the study is intended as a fresh contribution to the existing work on learning and the N400. In other respects, the present study has parallels to previous work on the evoked response, such as the mostly clinically oriented research using a picture-word paradigm (Stelmack and Miles 1990; Ford et al. 2001; Mathalon et al. 2002; Mathalon et al. 2010; Kuipers et al. 2018). The remaining sections describe and discuss the subject profile, data acquisition, experimental paradigm, noise reduction and the data analysis strategy. In doing so, they intend to follow as closely as possible the guidelines provided by Gross et al. (2013) for reporting MEG research. The Python and R code used to process and analyze the data after noise reduction are available at https://github.com/eeoluus/ma-thesis, together with replication instructions.

3.2.1 Subjects, their preparation and MEG measurement

Subjects, of whom thirty-two were included in the final analysis, were recruited by email, whereby they were thoroughly informed of the study and the course of the measurements. They were healthy, had a neurologically normal profile, reported no learning difficulties or psychiatric disorders, nor did they use drugs that affect the central nervous system. Participation was voluntary, and each subject received compensation in the form of a movie ticket or gift card at the end of the study. Written

consent was obtained from each. Subjects were pseudonymized and their data were treated as confidential. The research was approved by the Ethical Committee of the University of Jyväskylä and was part of the ChildBrain and PredictAble projects that were funded by the Department of Psychology and the European Union. See Xu et al. (2020) for a previous study of this sample.

Subjects were instructed to arrive without makeup, hair products, piercings or accessories. Upon arrival, they were given magnetically inactive clothes and shoes if needed, and magnetically active accessories were removed. This was done to minimize noise during the MEG recordings. The vision of the subjects was corrected to normal, if not normal, using non-magnetic eyeglasses available at the laboratory. Before measurements, they were familiarized with the MEG laboratory and equipment. Information was given about the structure of the measurement session. Two electrodes were placed diagonally around the eyes for electro-oculography (EOG). In order to correct for head movements, head position indicator (HPI) coils were placed behind each ear and on the forehead. Head shape and the placement of the coils were digitized using Polhemus Fastrak®, using the preauricular points and nasion as anatomical references. EOG and head digitization were done to enable further noise reduction during preprocessing of the data.

Subjects were sat under the gantry of the MEG device, and the seat was adjusted to reduce muscle tension in the neck such that the top of the head comfortably touched the helmet containing the sensors. The position of the gantry was upright at 68 degrees. The measurements took place in a magnetically shielded room. The MEG system used in the study was the 306-channel Elekta Neuromag® TRIUX. The sensor array of the TRIUX contains 102 magnetometers and 204 gradiometers that are stacked on top of each other to form 102 sensor triplets. A sampling frequency of 1000 Hz was used with a 0.1 Hz high-pass acquisition filter. The screen where stimuli were projected during the measurements was placed one meter in front of the subjects. Subjects were informed that they were to see a series of stimulus pairs together with a symbol telling whether the two stimuli belonged together or not. In another block, they were to see just the stimulus pairs and decide by a button press whether the stimuli belonged together. A response pad was given to the subjects for this purpose. Finally, subjects were instructed to stay as still as possible during the measurements.

3.2.2 Structure of the experiment

The experiment consisted of alternating learning and testing blocks. In the learning blocks, subjects were presented series of picture-word pairs that were followed by either of two symbols designating whether the picture and word belonged together or not. In the testing blocks, they were presented the same picture-word pairs, but this time a question mark followed. Upon seeing the question mark, subjects had to answer

whether the picture-word pairs were congruent or incongruent. After the testing blocks, they were also given automatic feedback on their performance on the screen. Both block types were introduced as practice blocks that were included to ensure that the subjects understood the task. Leaving out the practice blocks that were subsequently ignored, there were 24 blocks in total, half of which were learning blocks and the other half testing blocks. Each block contained eight congruent and the same number of incongruent trials. To characterize the development of behavioral and neurophysiological responses, the data from testing blocks was divided into early, early-middle, late-middle and late temporal bins. Only the testing blocks were included as active memory retrieval is more likely to happen during testing. The structure of the experiment is illustrated in Figure 3.

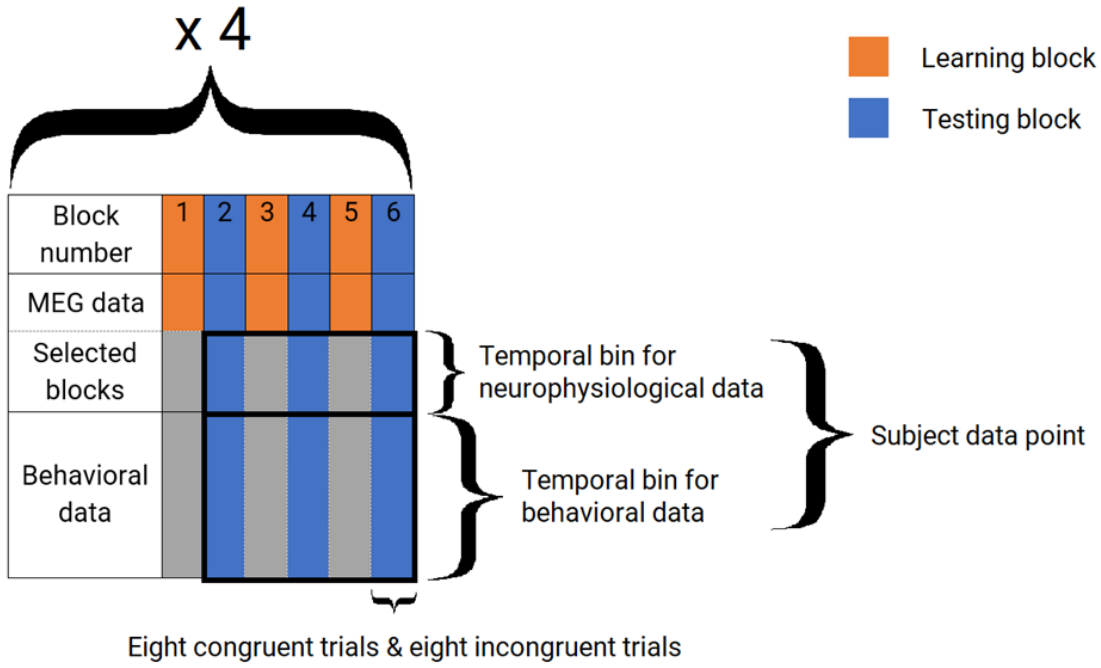

FIGURE 3 The structure of the experiment. For convenience, only six consecutive blocks are shown. The first row shows that learning and testing blocks alternated throughout the experiment. The second and third rows, separated by a dashed line, show that while MEG data was collected in every block, only the data collected during the course of every three consecutive testing blocks were entered in a temporal bin. The fourth row shows that behavioral data were collected only in testing blocks, and those from every three successive testing blocks were included in a temporal bin. The temporal bins are separated bolded lines, and the dashed lines within them indicate that their data were treated as wholes. Thus, evoked responses were obtained from averaging together epochs within each bin. Likewise, the corresponding behavioral measurements within each bin were averaged to provide measures of mean performance. Subject data

points were obtained by combining the size of the congruence effect with mean performance within each bin.

The structure of the trials was as follows. To orient their gaze, subjects were first shown a fixation cross for one second. This was followed by presentation of the picture for the same amount of time. Then, a fixation cross again appeared for one second. The word was subsequently shown for 1500 milliseconds. Finally, subjects were shown either a congruence or incongruence symbol for 1500 milliseconds in learning block trials, whereas in testing block trials a question mark was shown for the same amount of time. The question mark prompted subjects to answer by a button press whether the picture and word belonged together. The buttons of the response pad used for this purpose were counterbalanced. In the testing blocks, the next trial started after the subject had given their answer. A typical testing block trial is illustrated in Figure 4. To recapitulate, each temporal bin contained three times eight incongruent testing block trials and the same number of congruent testing block trials. For the evoked responses, there were therefore initially 24 trials to pick epochs from. This was deemed sufficient to provide adequate signal-to-noise ratio (SNR) while simultaneously imposing temporal bins.

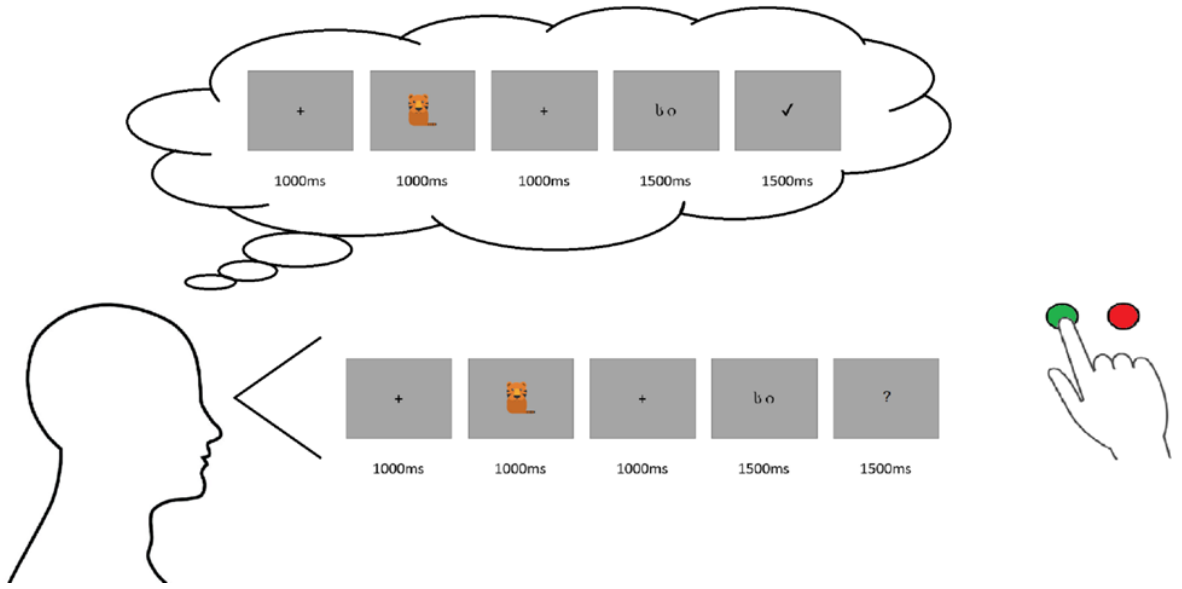

FIGURE 4 In testing blocks, subjects had to decide by a button press whether the picture-word stimulus pair was congruent or not. This illustration shows a subject engaged in active memory retrieval during testing. Included are also stimulus durations, which were not shown to the subjects. The thought bubble contains the related learning block trial, which differs only in that the question mark is replaced by a congruence symbol. In an incongruent learning block trial, the last stimulus would have been an incongruence symbol.

3.2.3 Noise reduction and evoked responses

The steps that followed data acquisition are procedures intended to improve SNR. Sources of noise in MEG include cardiac, muscular and ocular activity, respiration, and head movement. Non-physiological noise in the recordings might originate from inside or outside of the sensor array in the case of dental work, system noise from malfunctioning MEG channels or power line noise. First, bad channels were manually removed and interpolated using MaxFilter (version 3.0.17). Head movement compensation using data from HPI coils was performed with the same program. Finally, MaxFilter was used to do temporal signal space separation (tSSS) to reduce both head-internal and head-external noise. In tSSS, a spherical signal space is assumed such that noise external to it can be removed, together with correlated strong internal noise. This allows the researcher to remove noise from dental work, for example.

Power line noise (50 Hz) was then removed by low-pass filtering the data at 40 Hz, with a filter length of 10 seconds and a transition bandwidth of 0.5 Hz. A finite impulse response type of filter was used for this purpose. The filtering was done using an in-house graphical user interface (GUI) of MNE Python. MNE Python (Gramfort et al. 2013) itself is a Python-based toolkit for processing MEG data. Physiological noise that included eye artefacts and cardiac activity was removed by means of independent component analysis (ICA). The idea of ICA is to represent the data as a set of uncorrelated time series, the topography of which can be plotted. Cardiac activity so separated is easily recognizable and can subsequently be removed. Similarly, components containing ocular activity have a recognizable frontal topography and are identified by comparing them with time series from EOG. Noise reduction with ICA was performed using the same in-house GUI, with FastICA as the algorithm.

Evoked responses were obtained by averaging together epochs within each temporal bin, as described in the previous section. In averaging, normally distributed noise that is present will tend to cancel itself out, improving SNR as a function of the number of averaged epochs. The epochs started 200 milliseconds before the onset of the word stimulus and lasted for 1000 milliseconds after it. The onset of the word stimulus was defined by adding a trigger-to-stimulus delay to word trigger times. Triggers are essentially time stamps in the MEG data, but due to minute discrepancies between them and the actual stimulus presentation times, a trigger-to-stimulus delay must be taken into account. The N400 evoked response was quantified as the mean amplitude between 300 and 500 milliseconds post-stimulus. MNE Python was used for the averaging.

3.2.4 Strategic choices in looking for effects

Due to the research questions and a lack of MEG studies specifically using picture-word semantic judgment as the behavioral measure, several constraints are imposed on the analysis strategy. First, because only the N400 effect is of interest here, its source will not be localized in the present study. The origin of the effect is irrelevant to its predictive power. Further studies will be needed to localize it in the brain and possibly corroborate the existing empirical research on the subject (Lau et al. 2008). Sophisticated methods exist to address this issue that are useful especially when realistic head anatomy from MRI is co-registered (Dale and Sereno 1993). Second, the lack of studies using similar paradigms means that there are no strong hypotheses concerning the specific MEG sensors where an effect might be present. It was noted above that most of the research on the N400 has used word or sentence stimuli, and that studies combining linguistic and non-linguistic stimuli have been in the minority. Therefore, the researcher is faced with the problem of which sensors to test for statistically significant effects.

In such circumstances, one might ask why the tests should not be performed at every sensor. This would provide an unbiased perspective for deciding whether there is a statistically significant congruence effect. However, performing tests at all sensors greatly inflates the risk of committing a type I error, that is, falsely concluding that an effect is present when there really is none. This is known in statistics as the multiple comparisons problem (Pagano and Gauvreau 2018: 285–288). The essence of the problem is that the probability of committing at least one type I error increases multiplicatively with the number of tests. Normally, for a single test, this probability is controlled by deciding an arbitrary value such as 0.05 which reflects the probability that an observed effect was just noise. To illustrate, let us assume that there are 100 sensors, and a test is performed at each. The probability of correctly rejecting the null hypothesis that there is no effect at any of them is $(1 - 0.05)^{100}$, which equals 0.0059. Conversely, the probability that at least one test will be significant when the null hypothesis is true is 1 – 0.0059, which equals 0.9941. This is the experiment-wise probability of committing a type I error, and it is obviously far from the desired 0.05.

Consequently, the number of tests must be reduced, or the multiple comparisons have to be taken into account in some way. Reducing the number of tests is problematic not only because little research has been conducted using similar paradigms, but also because the number of sensors in modern MEG devices is even greater than the one hundred in the example above. The total number of sensors in the Elekta Neuromag® TRIUX MEG device that was used in the study is 306. Even when the 102 magnetometers are left out, there are still 204 gradiometers to provide adequate coverage of the head. Fortunately, procedures have been developed that correct for multiple comparisons. These allow for multiple tests to be performed while maintaining an

appropriate experiment-wise type I error rate. Depending on the procedure, the corrections might come with a cost of statistical power, referring to the ability of the tests to detect effects when the null hypothesis is false. However, permutation procedures have been demonstrated to retain a great degree of statistical power compared to some other methods, including different false discovery rate corrections (Fields and Kuperberg 2020). They emerge as one of the best options for dealing with multiple comparisons and therefore will be discussed in the following.

Permutation methods in statistics are grounded on the insight that, if the null hypothesis is true and there is no difference between population means, observations drawn from the populations are exchangeable. Blair and Karniski (1993) proposed a permutation procedure for a within-subjects design with two experimental conditions and multiple response variables that is applicable to the present data. In experimental settings where each subject contributes data from two experimental conditions, the data from one condition is exchanged with that from the other, within each subject. The different scenarios in which the data has been exchanged within some subjects but not the others make up the permutations for the entire data set. For each permutation, a t-statistic for each response variable is calculated. For example, if there are 100 sensors that measure at 100 time points for a single condition, 10 000 t-statistics will be calculated. In the method of Blair and Karniski, the largest of these termed t_{max} will be chosen to contribute to the null distribution. The null distribution refers to a distribution of all expectable t-statistics if the null hypothesis were true. If any test from the unpermuted data exceeds a critical value determined by a cut-off region from this distribution, it is deemed significant.

To show how the permutation procedure is applicable to the present neurophysiological data, its structure is briefly described. Every temporal bin contained the 300–500 milliseconds mean amplitudes from evoked responses to the word stimulus. Furthermore, magnetometer readings were disregarded because they do not measure as focally as gradiometers do. Therefore, the neurophysiological data of each subject were 816 mean amplitudes – from 204 gradiometers in four temporal bins – for each experimental condition. These constituted the multiple response variables. To determine the specific sensors and temporal bins where an effect might be present, the t_{max} procedure was applied. This yielded a null distribution from which a 0.05 cut-off region was specified. The critical value separating this cut-off region was used to rule out which of the original 816 t-statistics were significant. Finally, the sensor where the effect was strongest across time was chosen to contribute to the regression model that is described below.

3.2.5 Repeated measures and GLS estimation

To determine whether the semantic congruence effect predicts performance in congruence judgment and to describe this relationship require consideration of a possible correlation structure in the data. In the choice of the statistical method, the researcher should be aware of whether the observations are independent of each other, for if this criterion does not hold, methods such as ordinary least squares regression are not valid. The challenge faced by any study that utilizes multiple measurements from the same subjects is how to account for the possibility that the measurements within each subject are correlated, and are thus not independent. For example, 10 students might be randomly sampled from a population in a longitudinal study where tests are performed on them on three occasions, but due to individual differences, each student will respond idiosyncratically. Statistical methods exist for dealing with the non-independence of repeated measurements within subjects, which include repeated measures analysis of variance (ANOVA) with the Greenhouse-Geisser or Hyunh-Feldt correction, mixed effects models, and GLS estimation. All of these assume a correlation structure, which specifies the exact ways in which non-independence within subjects affects the outcome (Harrell 2015: 145).

In the mixed effects model, the problem of non-independence is solved by treating subjects as a random factor. Therefore, each subject exerts a random effect on the outcome. The effects are assumed to be random variables, each of which is normally distributed with a mean of zero and some specified variance. The researcher specifies the variances by choosing a correlation structure. In repeated measures ANOVA, the structure is assumed to be compound symmetry which, if not met, might require either the Greenhouse-Geisser or Hyunh-Feldt correction. A correlation structure is also chosen for the residuals, which are random in mixed effects models. For example, when evenly spaced observations close in time are assumed to exhibit higher degrees of correlation, the first-order autoregressive correlation structure might be preferred (West, Welch and Galecki 2014: 20–21). While in many ways more flexible than the repeated measures ANOVA with the appropriate corrections, some downsides of mixed effects models are that significance testing of the random effects is problematic and that they might not be normally distributed (Harrell 2015: 146).

In the present study, the non-independence of observations within subjects was accommodated by building a regression model using GLS. This permitted the inclusion of a continuous predictor variable more easily than an ANOVA would have. The N400 was included in the model as the sole predictor of performance. It was quantified for each individual and temporal bin as the difference between mean amplitudes during the 300–500 milliseconds post-stimulus time period at a sensor location determined by the initial, exploratory analysis. The development of performance over time was also analyzed using GLS estimation. The use of GLS results in a marginal model,

which differs from mixed effects models in its omission of random factors, although a correlation structure is still specified. In comparison to mixed models, the marginal model is a parsimonious yet reasonable choice when the objective is to obtain population-level estimates of effects (Pekár and Brabec 2016). The correlation structure was determined on the basis of values from the Akaike and Bayesian Information Criteria (AIC and BIC), which reflect the likelihood of model parameters penalized by their number.

In the context of mixed effects models, the significance testing of fixed effects requires a reference model and a null hypothesis model where the regression coefficient for the effect being tested is zero. The reference model is a full model that includes the hypothesized effect, and which is fit to the data. The correlation structures of the two models are the same, such that the only difference between them is the absence of the hypothesized effect in the null model. The test statistic is then derived from the likelihoods of the models, which follows a chi-square distribution with degrees of freedom equivalent to the difference in the number of parameters in the models (West et al. 2014: 35). A cut-off value is determined from this distribution and if the statistic exceeds a critical value, the effect is deemed significant. There are also other ways to test the significance of fixed effects which require only the reference model, but involve complex estimations of the degrees of freedom (West et al. 2014: 36–38). In the present study, the Wald statistic was used, which does not require fitting multiple models.

4 FINDINGS

This chapter describes the analytical procedure and the attained results. It is divided into two parts, the first of which concerns the MEG data. To start with, group-level differences between the experimental conditions are plotted and inspected. Following these first impressions, the condition differences in the N400 time window are tested in search of semantic congruence effects. Since no guiding hypothesis restricts the testing to specific sensors, permutation testing is performed to the whole sensor array. From this exploratory phase, the sensor with the strongest difference is picked to contribute to the data points. The second part constructs a regression model, where differences measured with this sensor predict performance in semantic congruence judgement. Full parameterization of the model is described, and the model is tested for statistical significance.

4.1 Semantic congruence effects

The data from gradiometers were averaged over subjects and trials within every period to depict group-level differences between the experimental conditions. Figures 5-8 show the resulting grand averages. There does not seem to be any distinguishable variation in the first period (Figure 5). In this and the following three figures, the color-coded lines show sensor-wise differences between the experimental conditions during the epoch in femtoteslas per centimeter (fT/cm). The colors correspond to the physical coordinates of selected sensors in the Elekta Neuromag® TRIUX array, shown schematically in the upper left corner. Note that since the sensor array is organized in triplets with sensors stacked on top of each other, the schematic shows only 102 locations. Due to their sensitivity to sources directly below them, only gradiometers were selected for plotting, which keeps the color-coding interpretable with respect to possible

underlying brain activity. Time runs on the horizontal axis, where zero marks the stimulus onset.

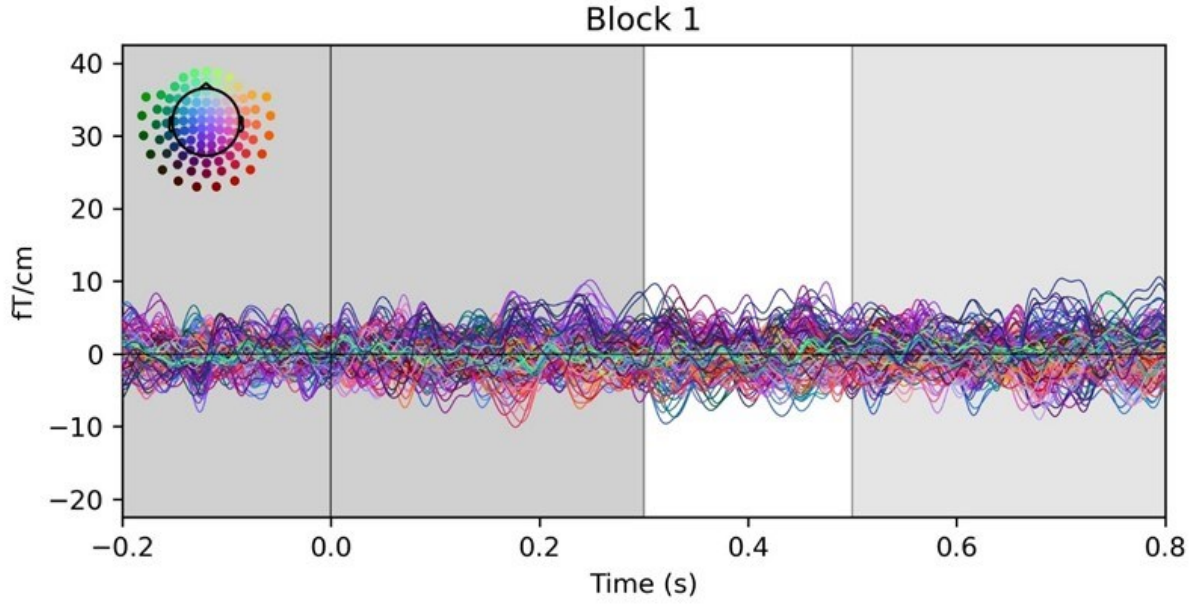

FIGURE 5 No distinguishable activity can be seen during the first period of the experiment.

Differences seem to be developing at central and left frontal locations in the second period (Figure 6). These correspond to the green, blue, and purple downward deflections in the selected time window. Moreover, a later central peak can be seen following the selected time window. In interpreting these observations, it should be kept in mind that gradiometers are most sensitive to magnetic fields produced by current dipoles oriented along specific axes. Depending on the direction of the dipoles, each gradiometer produces either positive or negative values. Thus, differences could in theory be produced by dipoles pointing to the opposite directions, or dipoles pointing to one direction that differ in strength. As information is lost in the subtraction, the plots do not provide any indication about the orientation of the individual dipoles produced by the two conditions: they only point to their difference.

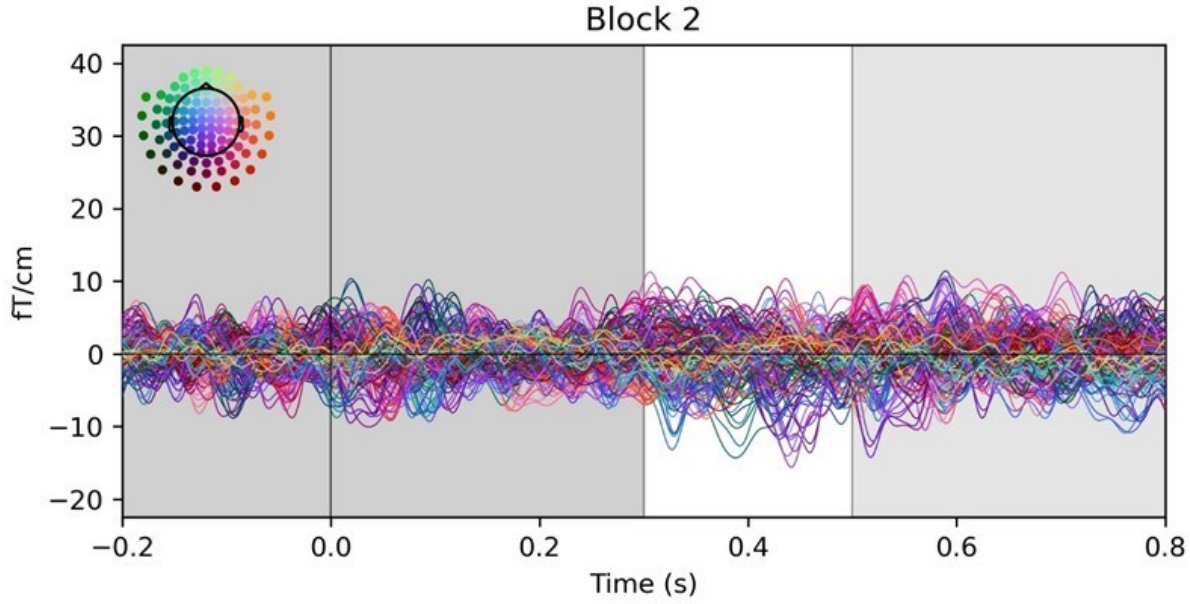

FIGURE 6 The plot shows how differences seem to be developing at central and left frontal locations. The late deflections occur at posterior locations.

In Figure 7, two almost simultaneous deflections appear at central gradiometers relatively late, preceded by seemingly differing activity at right occipital locations. Another strong difference is found at a left occipital location, which trails over the boundary of the time window. As in the second period, the most pronounced activity can be seen at posterior sensors, shown in purple. Here, the posterior activity is extended in time. At the end of the time window, the posterior negativity is accompanied by a simultaneous positive peak at more frontal sensors, shown in blue. There is also activity at right posterior sensors in the beginning of the time window.

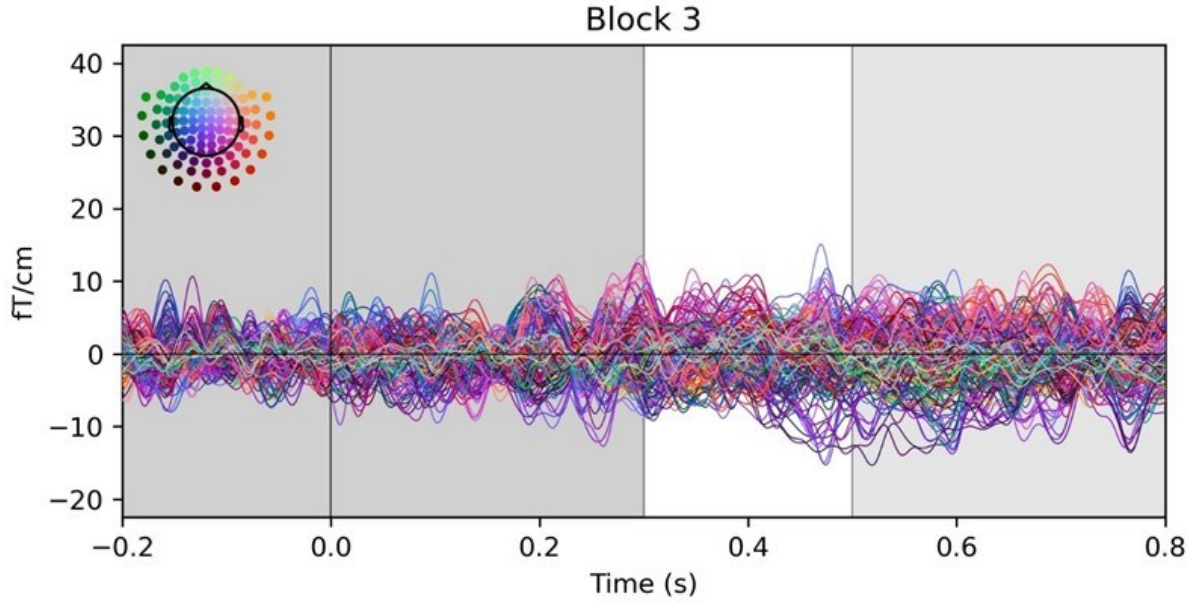

FIGURE 7 The posterior deflections are now extended in time. In contrast, right posterior sensors show activity in the beginning of the time window.

In Figure 8, the extended posterior activity seems to have developed into a response reaching its maximum amplitude right after the chosen time window. Together with the other posterior differences, this forms the most prominent peak in all of the plots. A similar, albeit weaker, peak occurs at roughly 600 milliseconds post-stimulus, which is accompanied by a positivity at slightly more frontal sensors. The central right activity, shown in pink, has also intensified in the beginning of the time window. All in all, it is apparent from this inspection that the classical N400 time window does not capture all of the variation in brain activity that might be of relevance to learning.

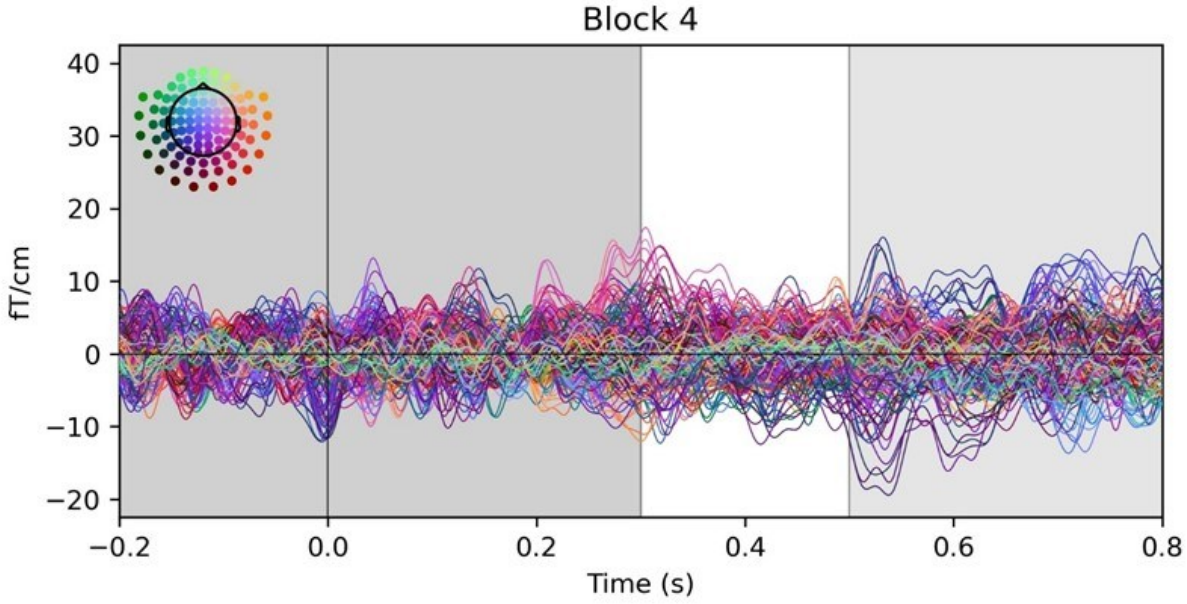

FIGURE 8 The previously described deflections have intensified, such that late posterior activity is distinguished from the plot with ease. Prominent deflections can be seen at the right central and posterior sensors too.

This visual, subjective inspection was followed by permutation-based testing for effects in order to provide a basis for the formation of data points. Mean amplitude differences in the N400 time window were subjected to the procedure to determine whether there were any significant differences from zero. In total, 1224 t-statistics were computed for this purpose, for the 306 sensors over the course of four periods. Sensors were not combined prior to this step. As the result of testing, none of the t-tests were significant at the level of $\alpha = 0.05$, which would have required a t-statistic surpassing 4.602. The t-statistics were used to produce the heatmap shown in Figure 9, which spans all sensors and periods. It was decided that data points would be formed on the basis of the sensor that came closest to reaching significance. Thus, the right occipital gradiometer MEG2542 was chosen to contribute to the data points ($t = 3.473$, $p = 0.591$, degrees of freedom = 31).

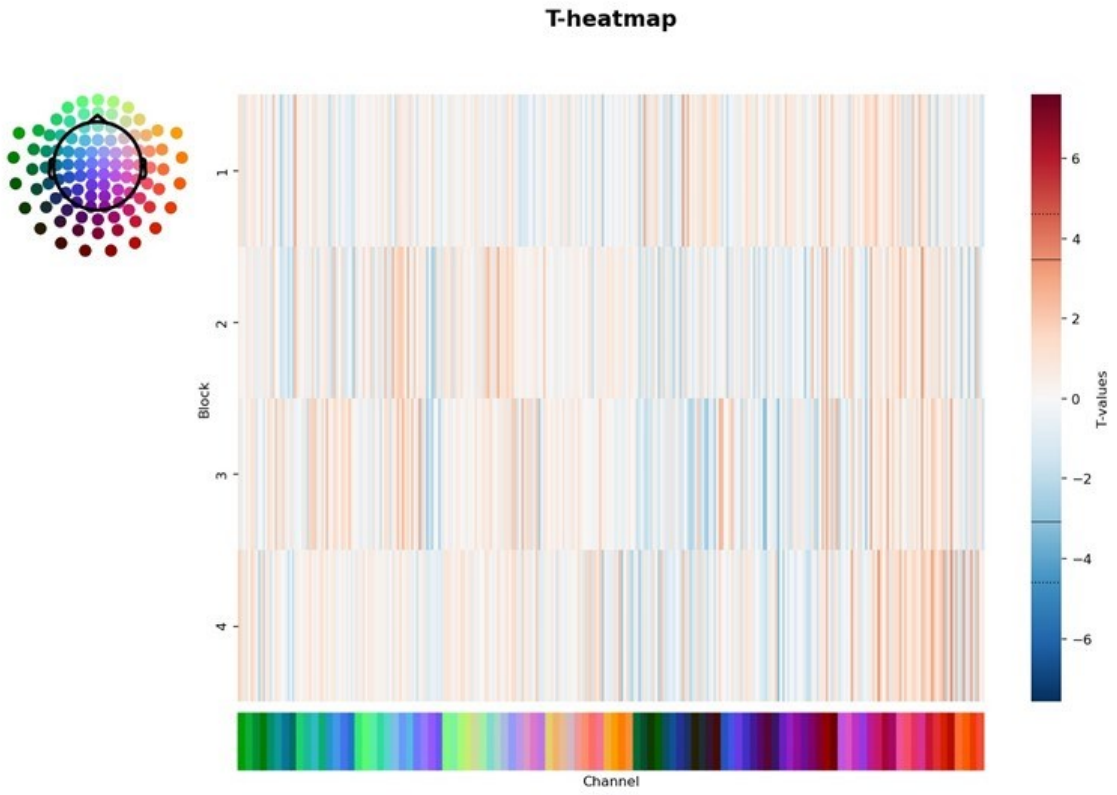

FIGURE 9 The t-statistics for all sensors and periods were plotted respectively as columns and rows of a heatmap. The color bar on the right shows how to interpret the colors of the map in terms of t-values. More positive t-values receive redder colors, while the more negative receive bluer. The highest observed t-values in either direction are marked with solid black lines on the bar, whereas the cut-off values that had to be surpassed for statistical significance are marked with dashed ones. In the bottom, the second color bar assigns sensors colors based on their physical coordinates, in a fashion similar to the previous figures. A schematic of the sensor array is shown in the upper left corner. Note that because magnetometers are now included as well, this color-coding mainly reflects the location of the sensors, rather than the dipoles generating the magnetic fields.

4.2 Building and testing a predictive model

The subsequent step was to form data points on the basis of brain activity and performance in the task. Initially, the purpose was to use accuracy as a measure of the latter, quantified as the percentage of correct responses in each period. However, it quickly became clear that many subjects had hit the performance ceiling during the experiment, rendering this measure invalid. Figure 10 depicts the participants' learning trajectories in terms of percentages. To circumvent the problem, new measures were

created as follows. Within each period, reaction times were measured from correct responses and averaged to produce a more valid measure of performance. Mean reaction times were regressed on block number using GLS, which yielded a strong effect ($t = -9.043$, $p < 0.001$, degrees of freedom = 126) indicating that learning had occurred. The use of reaction times mitigates the issue since they cannot humanly reach the floor of zero milliseconds. This decision is further discussed in the following chapter. Figure 11 depicts the participants' learning trajectories in terms of mean reaction times. Thus, data points were formed on the basis of mean reaction times and mean amplitude differences in the N400 time window at the right occipital gradiometer MEG2542.

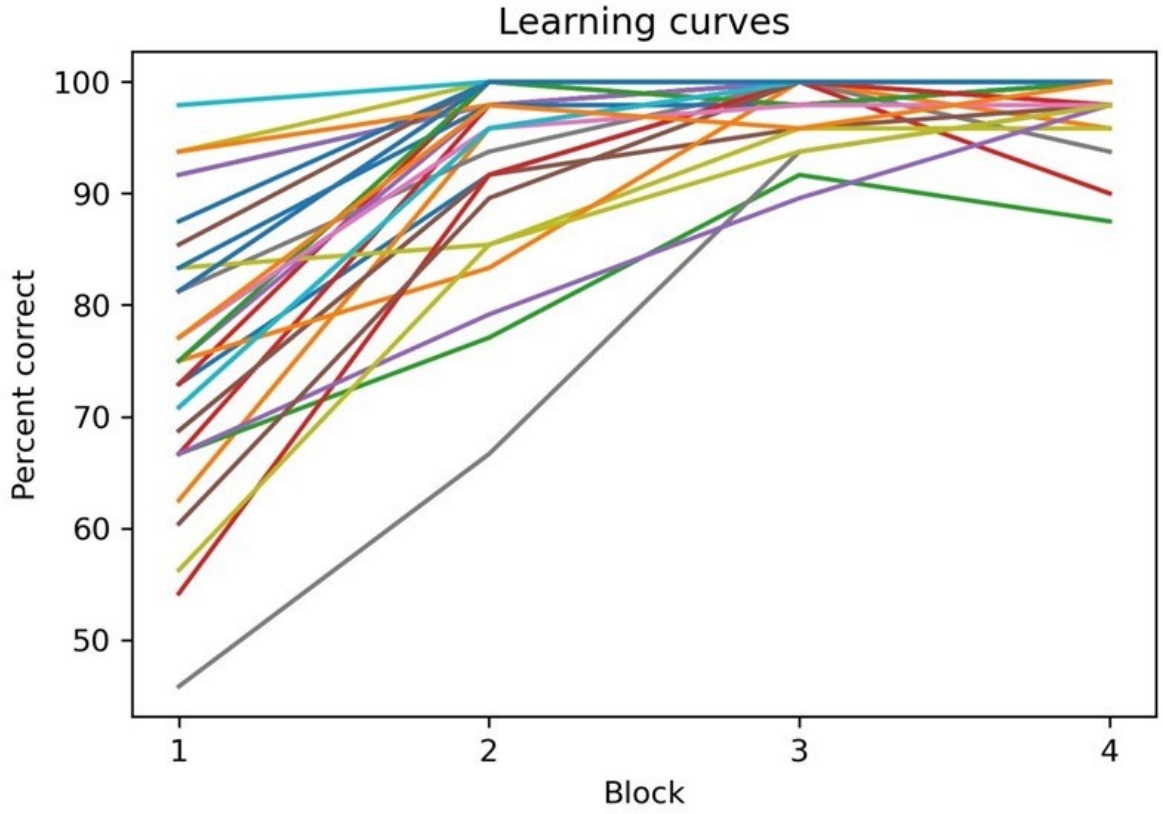

FIGURE 10 The colored lines show how the performance of individual participants has developed throughout the experiment. They suggest that participants have learned the associations between the words and the pictures, and that the training was effective. Both the ease of the task and the 100 percent ceiling imposed by quantifying in this way limit the variance in performance, which could be problematic.

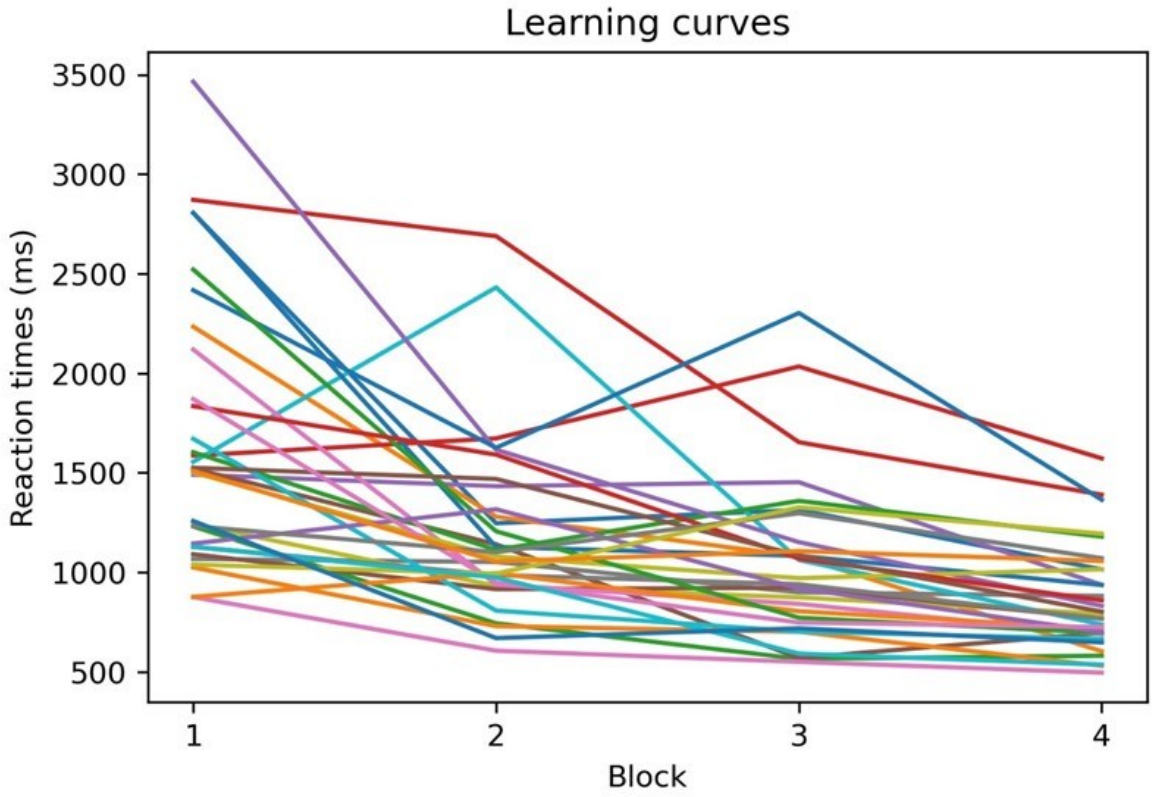

FIGURE 11 Learning curves in terms of reaction times. Again, the development of the performance of individual participants is shown, but this time in terms of reaction times. The reaction times were taken from correct responses. See Figure 6 for performance in terms of hit rate. In contrast to accuracy, this does not suffer from artificial ceiling or floor effects.

The data points were entered in a statistical model where the mean difference was used to predict learning outcomes. Each participant contributed four data points, and these repeated measures were taken into account by fitting a correlation structure within the participants. Having tried both compound symmetry and the first-order autoregressive structure, AIC and BIC values were in favor of the former (Table 1). Compound symmetry was also used in regressing reaction times on block number. In summary, the N400 mean amplitude difference was used as the sole predictor of mean reaction times in each period, having used compound symmetry to model within-participant correlation. For standardized and normalized data, the estimated β coefficient of the predictor was -0.11 with lower and upper boundaries of -0.24 and 0.02 as the lower and upper boundaries of the 95 percent confidence intervals. Using a Wald test, the effect did not reach significance ($t = -1.726$, $p = 0.087$, degrees of freedom = 126), but fared better than the N400 semantic congruence effect. The statistical model is depicted in Figure 12.

TABLE 1 The Akaike and Bayesian Information Criteria values for the two correlation structures, in which higher values are preferred. AIC and BIC were in agreement in favor of compound symmetry.

Model	Correlation structure	AIC	BIC
Performance as a function of the N400	Compound symmetry	1900.84	1912.186
	Autoregressive	1881.877	1893.223
Performance as a function of time	Compound symmetry	1901.066	1912.411
	Autoregressive	1892.955	1904.301

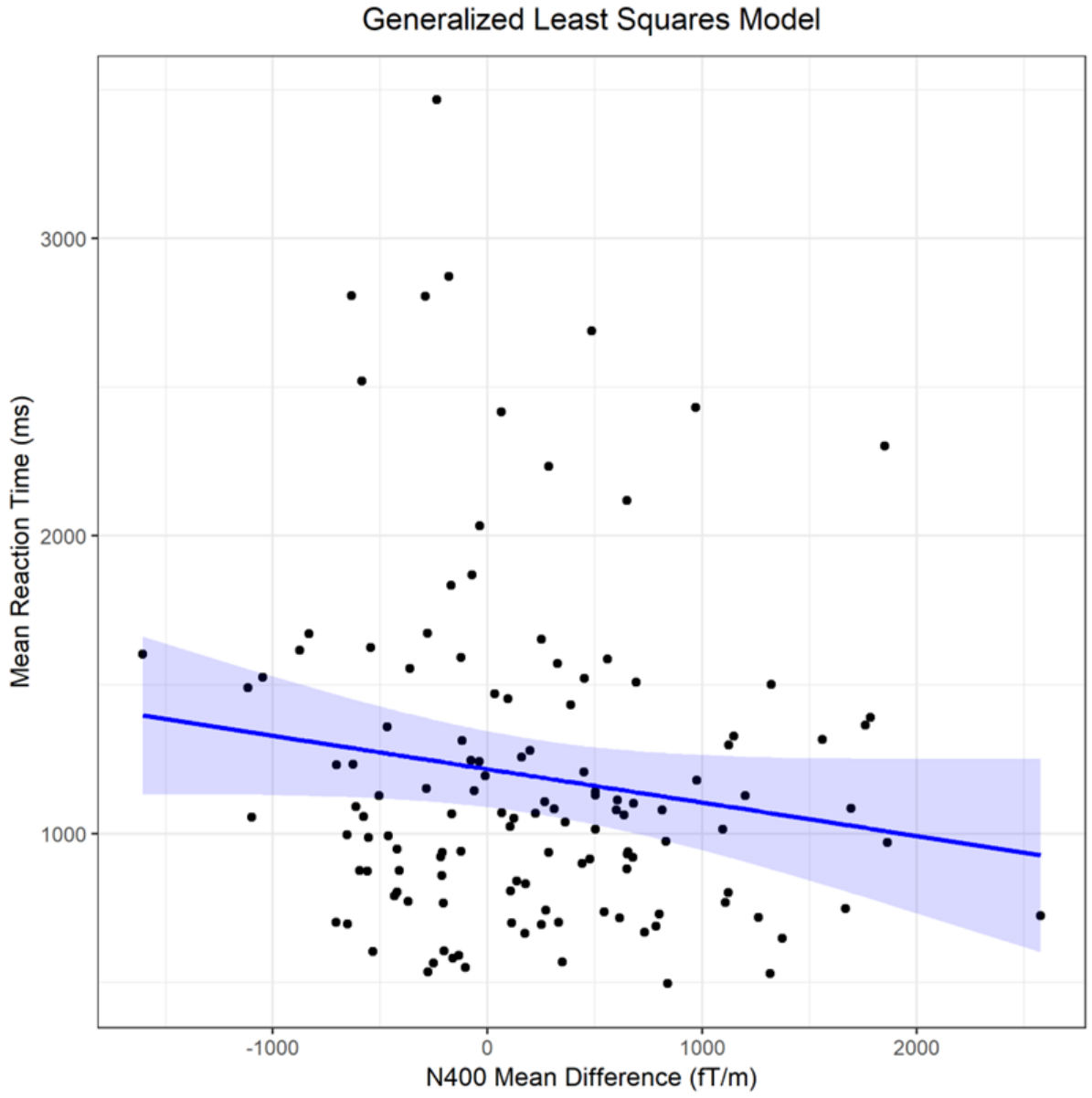

FIGURE 12 The statistical model. The negative slope indicates that more positive differences at the right occipital sensor are associated with shorter reaction times, that is, learning. Positive differences indicate that one of the conditions has received more positive values than the other. As noted earlier, this may be due a source dipole being stronger than the other or having reverse polarity. Regardless of

significance, the amount of dispersion seen in the plot indicates that much unexplained variance is left. The blue line shows the conditional mean of reaction times, given the N400 mean differences. The light blue area shows the 95 percent confidence interval for its estimated slope.

5 DISCUSSION

Differences in the N400 amplitude did not predict word learning, which is a novel result given the paucity of studies using similar regression models. This finding cautions against the use of the response as a biomarker of semantic learning, as defined in the Theoretical Background. Failing to discover significant amplitude differences through permutation-based testing in the exploratory phase was surprising, and not in line with most of the reviewed studies. Post hoc analyses were not performed (Kriegeskorte et al. 2010), although visual inspection suggested that amplitude differences were produced at a later time point. This would be consistent with the results of Ojima et al. (2005, 2011), where latency differences distinguished proficiency groups and developmental stages in vocabulary acquisition. Here, the encoding-retention delay was shorter than in the previous studies, some of which spanned days or months. In the very beginning of the learning process, effects may be found after the classical N400 time window. Paradoxically, while latency of the N400 is considered stable (Federmeier, Kutas and Dickson 2016), a methodological review found that highly variable time windows have been used to quantify the response (Šoškić et al. 2021). This practice is not optimal, as it can obscure important latency differences.

Methodology-wise there were both merits and shortcomings in the study. Most previous studies were designed with two outcomes in mind, the N400 and a task performance metric. In contrast, the present study tried to predict performance from the evoked response, which is a notable merit. From the point of view of education, to justify boosting the N400 either directly or indirectly, it must influence performance, and regression modelling is a more fitting strategy to determine this. The model, comprising the predictor and the outcome, could have been complemented with co-predictors, thus accounting for possible uncontrolled properties in the stimuli. Perfect experimental control is seldom possible, so it might be better to be explicit about confounding variables during model building (Sassenhagen and Alday 2016). Modelling was preceded by an exploratory analysis, where the entire sensor array was analyzed

in various phases of the experiment. The loss of statistical power due to the correction of multiple comparisons may have contributed to the lack of significant differences (Field and Kuperberg 2019). As an option, every time point could have been analyzed using a correction like threshold-free cluster enhancement (Smith and Nichols 2009). The sensors could have been combined to reduce the number of tests.

While the present work is mainly original research, a substantial part of its value lies in connecting work from distinct branches of research with N400 studies, ranging from neuropsychological and neurolinguistic work on the ATL and other central brain structures to systems consolidation. What received less attention were socio-cultural theories in linguistics, which criticize the idea that language is stored in the mind, in favor of its definition as a social practice in constant fluctuation. It should be noted here that the supporting role of the brain as a physical structure does not impose consensus on word meanings. In fact, that a speech community relies on memory representations of words to communicate does not make them immune to variation over time. There being memory representations of word meanings evidently did not stop semantic drift from happening to the common English word *nice*, previously designating foolish, nor other such developments (Traugott 2017).

Despite it being a socio-cognitive process, the sociality of language learning was not in focus. In the future, how semantic memory interacts with social cognition could form a fruitful branch of research. Moreover, if the meanings of words cannot be fully acquired, but are subject to social calibration and active construction, then their learning is an ongoing process. A short-term memory paradigm like the visual word learning paradigm that is described in the Methods provides but one entry point to this process. The present experiment relied on ostensive definitions, that is, ones pointing to examples, though it is difficult to see how a word like *solidarity* could be defined ostensibly. The idea that words possess meaning potential in place of discrete meanings (Hanks 2000) is explicit in the SG model itself, which the authors present (Rabovsky et al. 2018: 9) as sympathetic with the wider notion of predictive coding in the brain (Friston 2005). Unlike in the connectionist model, the "training" for humans is socially guided, lifelong, and has no predetermined end point.

Increasingly many N400 studies with educational undertones have been conducted, but there is possible conflict between the need to implement the curriculum and the interest to single out cognitive subprocesses. The functional profile of the N400 is still under refinement; for one, an account was presented in the Theoretical Background where it reflects the manipulation of semantic knowledge. On the contrary, the curriculum has a high-level impact on assessment. When performance does not occupy an unambiguous domain in the curriculum, the psychometric approach to assessment does not generalize well. Teasdale and Leung (2000) present the example of spoken English as an additional language from Welsh and English contexts, where

it is taught across boundaries of subjects. Interpreting how curricular goals are neurocognitively attained is possible only after establishing them. Arguably, neither semantic cognition nor word semantics occupy overt roles in language teaching, being subsumed under vocabulary knowledge together with stylistic understanding and morphosyntactic knowledge instead. Pedagogical and curricular considerations take precedence.

Finally, a main underlying theme has been the perspective of the educator and the practical challenges they face. Instruction through speech or writing is the most frequent educational intervention, which includes assigning tasks, giving feedback, and guiding the attention of the learners. Learning, the main outcome, is in principle inferred from behavior rather than the brain. If teachers require neuroscientific knowledge, it obviously differs from that appreciated by neurosurgeons. It was suggested in the Introduction that educational neuroscience could inform lifestyle choices which promote learning, and that practitioners could then spread awareness of them. This is not to downplay the possible effects of bodily activities such as handwriting or enacting word meanings, which lie outside the scope of the present work. Multidisciplinary collaboration is needed, and having a neuroscientific perspective could aid in making sense of results from sleep, nutrition or sports studies in relation to vocabulary learning. Arguably, although it cannot and should not replace behavioral work, the knowledge accumulating from the neuroscience of education surely complements such work. The following section further explores this theme in an attempt to flesh out lifestyle factors that contribute to learning.

5.1 Lifestyle of the good language learner

If the processes in the brain that underlie learning are known, the next question is what the educator should do with this information. As stated before, the principal value of neuroscience for education lies in its interdisciplinary nature. Therefore, it has the capability to unify findings from different fields of study, including but not limited to exercise, sleep, or nutritional studies, provided they are expressed in neuroscientific terms. The primary goal of educational neuroscience is to speed up the cerebral processes that account for learning, and progress in the aforementioned fields might yield useful educational interventions. This translative endeavor requires holistic understanding of how the processes in support of learning and their interventions are linked. This section ends the chapter with a multidisciplinary overview of findings and themes which might prove relevant to learning interventions. The discussion is admittedly cursory, but it provides some background for further

investigation of lifestyle factors. At the end, questions for future research are presented together with possible objections to such a project.

Sleep is such an important and broad topic in the context of learning that would require its own chapter. Here, the synaptic homeostasis hypothesis (SHY, Tononi and Cirelli 2014) suffices to illustrate proposed memory benefits of sleep. The proposal starts from the common view that synapses, the connections between neurons, are where memories are stored. According to SHY, waking hours increase the strength of these connections in the neocortex, as the individual learns about their environment. This initially produces memories of low signal-to-noise ratio, taking up a lot of capacity in the neocortex. Were the strength of synapses not normalized over time, the energy expenditure of neurons would grow too high. It is proposed that sleep does this by depressing the connections that code for irrelevant information, sparing those coding for statistical regularities. The learning capacity of the brain is thereby restored, which owes to spontaneous brain activity, primarily slow-wave activity. Tononi and Cirelli (2014) argue further that the direction of plastic changes could depend on the activity of neurotransmitters like noradrenaline, which is also associated with stress.

Stress is often thought to hinder learning, while some argue that it would be more productive to consider stress as an endogenous way to boost it (Rudland, Golding and Wilkinson 2020). Different hormones, including epinephrine and glucocorticoids such as cortisol, carry out the effects of stress in the body. The secretion of epinephrine is followed by a chain of activation that involves the vagus nerve and intervening nuclei in the brain. Glucocorticoids influence memory through a pathway called the hypothalamus-pituitary-adrenal axis (Smith and Vale 2006). Here, the hypothalamus processes stress-related signals and initiates an activation chain that culminates in the release of adrenocorticotropic hormone (ACTH). Glucocorticoids are released into circulation by the adrenal glands in response to ACTH. In both cases, noradrenergic activation of the amygdala is required for stress to promote learning (LaLumiere, McGaugh and McIntyre 2017). In rodents, the administration of ACTH has been found to increase performance up to a point, after which it decreases, but it is unclear whether this result applies to humans (Rudland et al. 2020).

Exercise is termed cardiovascular if it involves continuous, rhythmic movement of large muscle groups. Hillman, Erickson and Kramer (2008) reviewed the effect of cardiovascular exercise on cognition, and concluded that it may produce cognitive improvement, especially in the elderly. Exercise might have a protective effect against cognitive deterioration associated with aging; however, school-age children showed no benefit from exercise in memory tests. A later meta-analysis pooled studies into two groups, which involved either acute exercise or a long-term exercise program (Roig et al. 2013). Memory tests were classified as short-term or long-term tests, depending on whether retention followed encoding by more than two minutes. Possibly

owing to increased expression of the neurotrophin BDNF (Liu and Nusslock 2018), a major positive effect was observed for long-term memory, which was enhanced when encoding was accompanied by acute exercise. Some evidence indicates that exercise is most effectively performed at moderate intensity; too intense exercise prior to retention may disrupt task performance due to heightened arousal and fatigue (Roig et al. 2013; Liu and Nusslock 2018).

Appropriate nutrition could promote learning as well. As an example of an essential nutrient, docosahexaenoic acid (DHA), an omega-3 fatty acid, promotes transmembrane signaling and is important to the cell membranes of brain cells, 30% of whose phospholipid composition is comprised of DHA (Gómez-Pinilla 2008). As the human body is unable to produce it on its own, humans rely on dietary supply of DHA, garnered from sources such as salmon. According to a review by Gómez-Pinilla (2008), DHA has been found to contribute to one of the main forms of synaptic plasticity, long-term potentiation, which refers to persistent strengthening of synapses and is associated with learning and memory. Here, DHA stimulates signaling pathways that function to trap new receptors to the post-synaptic membrane, thus consolidating the synaptic changes (Rudy 2014). Finally, the review also considered antioxidant foods such as berries, which help alleviate oxidative stress that degrades neuronal membranes. Further investigation is needed to determine whether including DHA-rich foods and antioxidants in one's diet promotes vocabulary learning.

Another way cognition might be improved involves abuse of psychoactive substances and prescription drugs. Fond et al. (2015) review the cognitive effects of prescription drugs in the healthy. Normally, these drugs are used to treat a variety of disorders, ranging from attention deficit hyperactivity disorder (ADHD), narcolepsy and depression to Parkinson's and Alzheimer's disease (PD and AD). The most effective have been amphetamine-based compounds like Adderall, which contains dextroamphetamine and is used to treat ADHD. Dextroamphetamine (d-AMP), a mixture of amphetamine salts, has been shown to improve long-term declarative memory in healthy participants, exerting its effects by stimulating dopamine release and inhibiting its reuptake (Smith and Farah 2011). There is also evidence of Tolcapone, a drug used to treat PD, improving episodic memory by inhibiting catechol-O-methyltransferase, which breaks down dopamine and noradrenaline. However, non-medical use of these drugs, let alone promotion of their use by educators is ethically questionable, and their side effects could be counterproductive.

This concludes the short overview of lifestyle factors that might prove beneficial to learning. At the time of writing, almost fifty years have passed since Rubin (1975) formulated her conception of the good language learner, specifying learning strategies which successful learners seem to employ. This work could be complemented by studying whether they engage in cardiovascular exercise before learning; whether they eat

a DHA-rich diet; whether it is safe to say that they always sleep well, and the extent to which stress is beneficial to vocabulary acquisition. Whether such factors improve cognition or memory generally, or are beneficial for vocabulary acquisition in specific remains an open question. Better characterization of the brain systems that support vocabulary acquisition amidst the declarative memory system is needed. In applied research, it is clear that designing neuroscientifically inspired educational interventions requires experts from different fields, and is hardly a pursuit that an individual teacher could undertake. Thus, in practice, teachers could collaborate with professionals in school healthcare to find the best methods to help their learners.

As always, there is room for skepticism. Optimization of a healthy brain that receives enough nutrition and sleep might be unwarranted. In fact, for an organ that has resulted from millions of years of evolution, such interventions could be counterproductive, while those suffering from a condition like ADHD may benefit more (Smith and Farah 2011). The perceived cost-benefit ratio might contribute to the fact that the funding going to clinical research far surpasses the amount designated to studies of cognitive improvement in the healthy. Furthermore, there is very little current knowledge of how the formation of semantic memories, not to mention vocabulary acquisition, occurs in the CSC system. The role of the N400 is but one piece in the puzzle. More granular characterization of the key regions has been lacking until recently (Jung et al. 2022). The relationship of kainate glutamate receptors in the ATL (Palomero-Gallagher, Amunts and Zilles 2015) with GABAergic interneurons and synaptic plasticity (Valbuena and Lerma 2021) could offer an important clue. As noted previously, the inhibitory dynamics that these cells mediate is linked with semantic cognition (Jung et al. 2017).

6 CONCLUSION

While linguistic meaning relies on social coordination (van Lier 2004) and, fundamentally, trust (Wacewicz and Żywiczyński 2018) between people where it forms and transforms in a constant flux, its learning requires functional memory on the part of the individual. This work centered on how the brain supports vocabulary learning and took specific interest in the role of semantic memory. Focusing on a single evoked response, the N400, it first contributed to existing research by setting it in the context of the CSC framework, which draws upon imaging and clinical evidence, as well as theories of systems consolidation. In this view, the response is more associated with the control of semantic knowledge than with its storage, which resonates well with a prominent modelling study (Rabovsky et al. 2018). Systems consolidation theories offer a view into the dynamics of this knowledge, connect memory with theories operating at the cellular level such as SHY (Tononi and Cirelli 2014), and, in taking the hippocampus-neocortex dialogue into account, provide a more holistic perspective than CSC alone can.

The present study aimed to find out whether the amplitude of the N400 has the power to predict vocabulary learning, and so it was hypothesized that the effect would be produced as an epiphenomenon of word acquisition going on in the CSC system. No evidence was found to support this hypothesis. To investigate the question, the congruence paradigm permitted monitoring the development of two variables of interest: amplitude differences in the selected time window and performance in the task, from which learning is inferred. Starting from the common assumption that brain activity accounts for the change in behavior, the analytical approach modelled performance on the evoked response. Majority of the previous studies have tailored their analyses to answer the opposite question, the dependence of the response on performance. This owes possibly to the use of ANOVA procedures, which in contrast to regression methods do not allow for a single, continuous-valued predictor like the N400 mean difference. Moreover, the analysis should serve as an example that

common problems such as within-subject correlation can be handled in regression, too, with even more flexibility.

The present study differed in its approach in comparison to previous word learning N400 studies and serves as an opening as well as an invitation for future studies to adopt regression methods that allow for rigorous statistical inference, even when confounding variables are present (Sassenhagen and Alday 2016). Although the present study did not explore these possibilities, modelling allows for presentation of more complex stimuli (e.g., narrative speech) without a strict need for experimental control, as long as the complexity is specified in the model. Considering that experimental control can at times be very difficult, if not impossible to achieve, modelling offers opportunities even to move from controlled experiments to the use of naturalistic stimuli. One of the best examples comes from a Berkeley group who, using fMRI, had their participants listen to a podcast to uncover a "semantic atlas" in the brain (Huth et al. 2016). In general, neuroscientists have become increasingly interested in what goes on "in the real world" (Matusz et al. 2019), which is undoubtedly welcomed by linguists who stress ecological validity.

In fact, an effect in the classical N400 time window is also elicited by natural speech presented in audiobooks, when it is calculated as a function of the semantic dissimilarity of the words to their contexts (Broderick et al. 2018). These temporal response functions (TRFs, Crosse et al. 2016) offer an alternative to evoked responses, since they permit the use of uncontrolled, complex and continuous stimuli. It is worth noting that TRFs, or encoding models in general, are statistically far more complex than the straightforward models used in the present work (Naselaris et al. 2011; Holdgraf et al. 2017). Given the recent interest in naturalistic study designs, a future study could substitute TRFs for evoked responses, making use of naturalistic stimuli, while including some performance metric to be predicted. As it happens, precisely such a study has recently been conducted (Ihara et al. 2021). Notably, the group seems to adopt an analytical approach similar to that of the present study, in which proficiency is predicted from brain responses.

In spite of the methodological merits, the results presented herein provide no evidence for the claim that amplitude differences in the classic N400 time window predict semantic learning. Visual analysis of the evoked responses suggests that latency of the response might be the better predictor, but this is for future studies to decide. Different time windows have been used to quantify the N400 (Šoškić et al. 2021), and conflating the effects might risk ignoring important details of how learning is implemented in the brain. Due to and despite the null results, the major focus in the discussion was the implications of positive results for pedagogical practice, were they obtained. For example, it is worth thinking about what educators are to make of such information, or how they could use it to design better teaching methods. After all, the

brain operates in the background, not only in the sense that sensory experience and activity of the learner depend on it, but also in the sense that it is not the most tangible asset for the teacher to work with.

If learning outcomes could be predicted from the N400 response, the processes that contribute to its generation could be subjected to manipulation in hopes of speeding them up, thus accelerating learning. The response itself may be a side effect of learning rather than its prerequisite, but could still indirectly reflect learning, with the activity of the control regions of the CSC system changing. While the previous chapter offered a sketch of some interventions under the theme of lifestyle, there could be a number of possible interventions. Better understanding of CSC processes, especially ATL physiology in systems consolidation, will be essential in their exclusion and selection. It is important to consider this in relation to what teachers do in practice. For one, scaffolding in a classroom may include directing the attention of the learners, controlling their frustration, simplifying task instructions, clarifying the goals, and pointing out differences between the current level of performance and the desired outcomes (van Lier 2004: 149–152). The physiology of word acquisition may seem quite distant in relation to this.

As with any applied research, the context of application poses practical constraints. What is practically possible is constrained by the expertise of teachers and the tools and opportunities available to them. For example, the school environment provides teachers with a world of (digital) books, projectors, conversation and paperwork, not one of pharmacological agents or brain imaging devices. Whether they are in a position to promote diets or exercise for learning, or to speak for the memory benefits of sleep, is also a question of professional boundaries. Unless their education is diversified, it would seem questionable to expect teachers to be healthcare professionals, which the preceding overview of topics like stress, drugs and sleep would envision. These issues remain, even if the cerebral processes that subserve learning could be pinpointed or corresponded to those that produce variation in the N400, were it found to predict task performance. Therefore, the processes need to be characterized in a way that it becomes possible to manipulate them in some practical way to speed them up, whereas the scientific pursuit for ever more granular understanding comes second.

Another point of criticism is that interventions like telling to sleep well or eat healthy are "no-brainers", in a very literal sense. From this point of view, progresses in educational neuroscience are of little novelty value. In retrospect, the tips and insights originating from the field may sound obvious, although innovations may spawn in the future with studies investigating questions like whether slow-wave activity could be boosted to improve learning. Some of the possibilities will undoubtedly involve high-level technology, while the other, "low-tech" interventions involve methods that are more accessible to the learners. In this way, it may be the mundane,

"no-brainer" interventions which best lend themselves to implementation in practice. Topics outlined in the previous chapter like diet, exercise or sleep involve elements of life which the learners have some control over. These would be the ingredients of a learning-friendly lifestyle. Although teachers are not healthcare professionals, they could participate in raising awareness of habits that promote or hinder learning. This could ideally take place in collaboration with school healthcare.

To sum it up, this work both explored the promises and possible pitfalls of the neuroscience of education and conducted a study in the field. In studying the connection between word learning and the N400, it contributed to existing work on the topic methodologically and interpreted it in terms of the theories of CSC and systems consolidation. While the amplitude of the response did not significantly predict learning outcomes, future studies should explore the predictive capability of its latency. Neuroscience has occupied a spot alongside many fields of science in the form of multidisciplinary collaboration, education being one of the most recent. This coincides with recent trends in neuroscience like the move towards naturalistic paradigms, which presents an opportunity also for education-oriented studies. As the brain operates in the background in all aspects of life, developments in the neuroscience of education may offer some yet unseen benefits to the learner. Lifestyle interventions could form one example, perhaps in collaboration with school healthcare. While knowledge of the learning brain is valuable in itself, utilizing it in pedagogical practice requires, to some degree, rethinking the profession.

REFERENCES

Anderson, P. W. (1972). More is different: broken symmetry and the nature of the hierarchical structure of science. *Science* [online] 177 (4047), 393–396. https://doi.org/10.1126/science.177.4047.393

Aronson, J. K. (2005). Biomarkers and surrogate endpoints. *British Journal of Clinical Pharmacology* [online] 59 (5), 491–494. https://doi.org/10.1111/j.1365-2125.2005.02435.x

Baker, S. G. and Kramer, B. S. (2013). Surrogate endpoint analysis: an exercise in extrapolation. *Journal of the National Cancer Institute* [online] 105 (5), 316–320. https://doi.org/10.1093/jnci/djs527

Baker, S. G. and Kramer, B. S. (2020). Simple methods for evaluating 4 types of biomarkers: surrogate endpoint, prognostic, predictive, and cancer screening.

Biomarker Insights [online] 15, n. pag. https://doi.org/10.1177/1177271920946715
Batterink, L. and Neville, H. (2011). Implicit and explicit mechanisms of word learning in a narrative context: an event-related potential study. *Journal of Cognitive Neuroscience* [online] 23 (11), 3181–3196. https://doi.org/10.1162/jocn_a_00013
Bevilacqua, D., Davidesco, I., Wan, L., Chaloner, K., Rowland, J., Ding, M., Poeppel, D. and Dikker, S. (2019). Brain-to-brain synchrony and learning outcomes vary by student–teacher dynamics: evidence from a real-world classroom electroencephalography study. *Journal of Cognitive Neuroscience* [online] 31 (3), 401–411. https://doi.org/10.1162/jocn_a_01274
Binder, J. R. and Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences* [online] 15 (11), 527–536. https://doi.org/10.1016/j.tics.2011.10.001
Blair, R. C. and Karniski, W. (1993). An alternative method for significance testing of waveform difference potentials. *Psychophysiology* [online] 30 (5), 518–524. https://doi.org/10.1111/j.1469-8986.1993.tb02075.x
Bowers, J. S. (2016). The practical and principled problems with educational neuroscience. *Psychological Review* [online] 123 (5), 600–612. https://doi.org/10.1037/rev0000025
Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J. and Lalor, E. C. (2018). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology* [online] 28 (5), 803–809. https://doi.org/10.1016/j.cub.2018.01.080
Carter, R. (2012). *Vocabulary: applied linguistic perspectives*. New York: Routledge.
Coch, D. and Benoit, C. (2015). N400 event-related potential and standardized measures of reading in late elementary school children: correlated or independent? *Mind, Brain, and Education* [online] 9 (3), 145–153. https://doi.org/10.1111/mbe.12083
Cocquyt, E. M., Lanckmans, E., van Mierlo, P., Duyck, W., Szmalec, A., Santens, P. and De Letter, M. (2020). The white matter architecture underlying semantic processing: a systematic review. *Neuropsychologia* [online] 136, n. pag. https://doi.org/10.1016/j.neuropsychologia.2019.107182
Cohen, N. J. and Squire, L. R. (1980). Preserved learning and retention of pattern-analyzing skill in amnesia: dissociation of knowing how and knowing that. *Science* [online] 210 (4466), 207–210. https://doi.org/10.1126/science.7414331
Cornejo, C. (2004). Who says what the words say? The problem of linguistic meaning in psychology. *Theory & Psychology* [online] 14 (1), 5–28. https://doi.org/10.1177/0959354304040196
Crosse, M. J., Di Liberto, G. M., Bednar, A. and 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* [online] 10, n. pag. https://doi.org/10.3389/fnhum.2016.00604
Crystal, D. (2004). The past, present, and future of World English. In A. Gardt and B. Hüppauf (eds.), *Globalization and the future of German*. Berlin: De Gruyter, 27–45.

Dale, A. M. and Sereno, M. I. (1993). Improved localization of cortical activity by combining EEG and MEG with MRI cortical surface reconstruction: a linear approach. *Journal of Cognitive Neuroscience* [online] 5 (2), 162–176. https://doi.org/10.1162/jocn.1993.5.2.162

Duffau, H., Moritz-Gasser, S. and Mandonnet, E. (2014). A re-examination of neural basis of language processing: proposal of a dynamic hodotopical model from data provided by brain stimulation mapping during picture naming. *Brain and Language* [online] 131, 1–10. https://doi.org/10.1016/j.bandl.2013.05.011

Duffau, H. (2015). Stimulation mapping of white matter tracts to study brain functional connectivity. *Nature Reviews Neurology* [online] 11 (5), 255–265. https://doi.org/10.1038/nrneurol.2015.51

Dufva, H., Suni, M., Aro, M. and Salo, O.-P. (2011). Languages as objects of learning: language learning as a case of multilingualism. *Apples - Journal of Applied Language Studies* [online] 5 (1), 109–124. https://apples.journal.fi/article/view/97818

Elgort, I., Perfetti, C. A., Rickles, B. and Stafura, J. Z. (2015). Contextual learning of L2 word meanings: second language proficiency modulates behavioural and event-related brain potential (ERP) indicators of learning. *Language, Cognition and Neuroscience* [online] 30 (5), 506–528. https://doi.org/10.1080/23273798.2014.942673

Federmeier, K. D., Kutas, M. and Dickson, D. S. (2016). A common neural progression to meaning in about a third of a second. In G. Hickok and S. Small (eds.), *Neurobiology of language*. San Diego, CA: Academic Press, 557–567.

Feiler, J. B. and Stabio, M. E. (2018). Three pillars of educational neuroscience from three decades of literature. *Trends in Neuroscience and Education* [online] 13, 17–25. https://doi.org/10.1016/j.tine.2018.11.001

Fields, E. C. and Kuperberg, G. R. (2020). Having your cake and eating it too: flexibility and power with mass univariate statistics for ERP data. *Psychophysiology* [online] 57 (2). https://doi.org/10.1111/psyp.13468

Fond, G., Micoulaud-Franchi, J.-A., Brunel, L., Macgregor, A., Miot, S., Lopez, R., Richieri, R., Abbar, M., Lancon, C. and Repantis, D. (2015). Innovative mechanisms of action for pharmaceutical cognitive enhancement: a systematic review. *Psychiatry Research* [online] 229 (1-2), 12–20. https://doi.org/10.1016/j.psychres.2015.07.006

Ford, J. M., Askari, N., Gabrieli, J., Mathalon, D. H., Tinklenberg, J., Menon, V. and Yesavage J. (2001). Event-related brain potential evidence of spared knowledge in Alzheimer's disease. *Psychology and Aging* [online] 16 (1), 161–176. https://doi.org/10.1037/0882-7974.16.1.161

Frey, B. B., Schmitt, V. L. and Allen, J. P. (2012). Defining authentic classroom assessment. *Practical Assessment, Research, and Evaluation* [online] 17 (1), n. pag. https://doi.org/10.7275/sxbs-0829

Friederici, A. D., von Cramon, D. Y. and Kotz, S. A. (1999). Language related brain potentials in patients with cortical and subcortical left hemisphere lesions. *Brain* [online] 122 (6), 1033–1047. https://doi.org/10.1093/brain/122.6.1033

Friston, K. (2005). A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences* [online] 360 (1456), 815–836. https://doi.org/10.1098/rstb.2005.1622
Ganis, G. and Kutas, M. (2003). An electrophysiological study of scene effects on object identification. *Cognitive Brain Research* [online] 16 (2), 123–144. https://doi.org/10.1016/S0926-6410(02)00244-6
Gómez-Pinilla, F. (2008). Brain foods: the effects of nutrients on brain function. *Nature Reviews Neuroscience* [online] 9 (7), 568–578. https://doi.org/10.1038/nrn2421
Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L. and Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience* [online] 7 (267), 1–13. https://doi.org/10.3389/fnins.2013.00267
Grieder, M., Crinelli, R. M., Jann, K., Federspiel, A., Wirth, M., Koenig, T., Stein, M., Wahlund, L.-O. and Dierks, T. (2013). Correlation between topographic N400 anomalies and reduced cerebral blood flow in the anterior temporal lobes of patients with dementia. *Journal of Alzheimer's Disease* [online] 36 (4), 711–731. https://doi.org/10.3233/JAD-121690
Gross, J., Baillet, S., Barnes, G. R., Henson, R. N., Hillebrand, A., Jensen, O., Jerbi, K., Litvak, V., Maess, B., Oostenveld, R., Parkkonen, L., Taylor, J. R., van Wassenhove, V., Wibral, M. and Schoffelen, J. M. (2013). Good practice for conducting and reporting MEG research. *NeuroImage* [online] 65 (15), 349–363. https://doi.org/10.1016/j.neuroimage.2012.10.001
Gräff, J. and Tsai, L. H. (2013). The potential of HDAC inhibitors as cognitive enhancers. *Annual Review of Pharmacology and Toxicology* [online] 53, 311–330. https://doi.org/10.1146/annurev-pharmtox-011112-140216
Hanks, P. (2000). Do word meanings exist? *Computers and the Humanities* [online] 34 (1/2), 205–215. http://www.jstor.org/stable/30204810
Hari, R. and Puce, A. (2017). *MEG-EEG primer*. New York: Oxford University Press.
Harrell, F. E. (2015). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis* (2nd edition). New York: Springer.
Hillman, C. H., Erickson, K. I. and Kramer, A. F. (2008). Be smart, exercise your heart: exercise effects on brain and cognition. *Nature Reviews Neuroscience* [online] 9 (1), 58–65. https://doi.org/10.1038/nrn2298
Hockett, C. F. (1958). *A course in modern linguistics*. New York: Macmillan.
Holcomb, P. J. (1993). Semantic priming and stimulus degradation: implications for the role of the N400 in language processing. *Psychophysiology* [online] 30 (1), 47–61. https://doi.org/10.1111/j.1469-8986.1993.tb03204.x
Holdgraf, C. R., Rieger, J. W., Micheli, C., Martin, S., Knight, R. T. and Theunissen, F. E. (2017). Encoding and decoding models in cognitive electrophysiology. *Frontiers in Systems Neuroscience* [online] 11, n. pag. https://doi.org/10.3389/fnsys.2017.00061
Holmes, J. (2013). *An introduction to sociolinguistics* (4th edition). London: Routledge.
Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E. and Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral

cortex. *Nature*, [online] 532 (7600), 453–458. https://doi.org/10.1038/nature17637
Ihara, A. S., Matsumoto, A., Ojima, S., Katayama, J. I., Nakamura, K., Yokota, Y., Watanabe, H. and Naruse, Y. (2021). Prediction of second language proficiency based on electroencephalographic signals measured while listening to natural speech. *Frontiers in Human Neuroscience*, 15. https://doi.org/10.3389/fnhum.2021.665809
Indefrey, P. and Davidson, D. J. (2009). Second language acquisition. In *Encyclopedia of neuroscience*, 517–523.
Jackson, R. L., Lambon Ralph, M. A. and Pobric, G. (2015). The timing of anterior temporal lobe involvement in semantic processing. *Journal of Cognitive Neuroscience* [online] 27 (7), 1388–1396. https://doi.org/10.1162/jocn_a_00788
Jung, J., Williams, S. R., Sanaei Nezhad, F. and Lambon Ralph, M. A. (2022). Neurochemical profiles of the anterior temporal lobe predict response of repetitive transcranial magnetic stimulation on semantic processing. *NeuroImage* [online] 258, n. pag. https://doi.org/10.1016/j.neuroimage.2022.119386
Jung, J., Williams, S. R., Sanaei Nezhad, F. and Lambon Ralph, M. A. (2017). GABA concentrations in the anterior temporal lobe predict human semantic processing. *Scientific Reports* [online] 7 (1), 1–9. https://doi.org/10.1038/s41598-017-15981-7
Klinzing, J. G., Niethard, N. and Born, J. (2019). Mechanisms of systems memory consolidation during sleep. *Nature Neuroscience* [online] 22 (10), 1598–1610. https://doi.org/10.1038/s41593-019-0467-3
Kotz, S. A., Opitz, B. and Friederici, A. D. (2007). ERP effects of meaningful and non-meaningful sound processing in anterior temporal patients. *Restorative Neurology and Neuroscience* [online] 25 (3-4), 273–284. https://content.iospress.com/articles/restorative-neurology-and-neuroscience/rnn253410
Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A. and Poeppel, D. (2017). Neuroscience needs behavior: correcting a reductionist bias. *Neuron* [online] 93 (3), 480–490. https://doi.org/10.1016/j.neuron.2016.12.041
Kriegeskorte, N., Lindquist, M. A., Nichols, T. E., Poldrack, R. A. and Vul, E. (2010). Everything you never wanted to know about circular analysis, but were afraid to ask. *Journal of Cerebral Blood Flow & Metabolism* [online] 30 (9), 1551–1557. https://doi.org/10.1038/jcbfm.2010.86
Krokfors, L. (2017). Opetussuunnitelman pedagogiset mahdollisuudet – opettajat uuden edessä. In T. Autio, L. Hakala and T. Kujala (eds.), *Opetussuunnitelmatutkimus: keskustelunavauksia suomalaiseen kouluun ja opettajankoulutukseen.* Tampere: Tampere University Press, 247–266.
Kuipers, J. R., Jones, M. W. and Thierry, G. (2018). Abstract images and words can convey the same meaning. *Scientific Reports* [online] 8 (1), 1–6. https://doi.org/10.1038/s41598-018-25441-5
Kumaran, D., Hassabis, D. and McClelland, J. L. (2016). What learning systems do intelligent agents need? Complementary learning systems theory updated.

Trends in Cognitive Sciences, [online] 20 (7), 512–534. https://doi.org/10.1016/j.tics.2016.05.004
Kutas, M. (1993). In the company of other words: electrophysiological evidence for single-word and sentence context effects. *Language and Cognitive Processes* [online] 8 (4), 533–572. https://doi.org/10.1080/01690969308407587
Kutas, M. and Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology* [online] 62, 621–647. https://doi.org/10.1146/annurev.psych.093008.131123
Kutas, M. and Hillyard, S. A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science* [online] 207 (4427), 203–205. https://doi.org/10.1126/science.7350657
LaLumiere, R. T., McGaugh, J. L. and McIntyre, C. K. (2017). Emotional modulation of learning and memory: pharmacological implications. *Pharmacological Reviews* [online] 69 (3), 236–255. https://doi.org/10.1124/pr.116.013474
Lambon Ralph, M. A., Jefferies, E., Patterson, K. and Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience* [online] 18 (1), 42–55. https://doi.org/10.1038/nrn.2016.150
Lau, E. F., Phillips, C. and Poeppel, D. (2008). A cortical network for semantics: (de)constructing the N400. *Nature Reviews Neuroscience* [online] 9 (12), 920–933. https://doi.org/10.1038/nrn2532
Liu, P. Z. and Nusslock, R. (2018). Exercise-mediated neurogenesis in the hippocampus via BDNF. *Frontiers in Neuroscience* [online] 12, n. pag. https://doi.org/10.3389/fnins.2018.00052
Luck, S. J. (2014). *An introduction to the event-related potential technique*. Cambridge, MA: MIT Press.
Marr, D. (1982). *Vision: a computational investigation into the human representation and processing of visual information*. San Francisco, CA: W. H. Freeman.
Mathalon, D. H., Faustman, W. O. and Ford, J. M. (2002). N400 and automatic semantic processing abnormalities in patients with schizophrenia. *Archives of General Psychiatry* [online] 59 (7), 641–648. https://doi.org/10.1001/archpsyc.59.7.641
Mathalon, D. H., Roach, B. J. and Ford, J. M. (2010). Automatic semantic priming abnormalities in schizophrenia. *International Journal of Psychophysiology* [online] 75 (2), 157–166. https://doi.org/10.1016/j.ijpsycho.2009.12.003
Matusz, P. J., Dikker, S., Huth, A. G. and Perrodin, C. (2019). Are we ready for real-world neuroscience? *Journal of Cognitive Neuroscience* [online] 31 (3), 327–338. https://doi.org/10.1162/jocn_e_01276
McClelland, J. L., McNaughton, B. L. and O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, [online] 102 (3), 419–457. https://doi.org/10.1037/0033-295X.102.3.419

McLaughlin, J., Osterhout, L. and Kim, A. (2004). Neural correlates of second-language word learning: minimal instruction produces rapid change. *Nature Neuroscience* [online] 7 (7), 703–704. https://doi.org/10.1038/nn1264

Naselaris, T., Kay, K. N., Nishimoto, S. and Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage* [online] 56 (2), 400–410. https://doi.org/10.1016/j.neuroimage.2010.07.073

Nozari, N. and Thompson-Schill, S. L. (2016). Left ventrolateral prefrontal cortex in processing of words and sentences. In G. Hickok and S. Small (eds.), *Neurobiology of language*. San Diego, CA: Academic Press, 569–584.

Ojima, S., Nakamura, N., Matsuba-Kurita, H., Hoshino, T. and Hagiwara, H. (2011). Neural correlates of foreign-language learning in childhood: a 3-year longitudinal ERP study. *Journal of Cognitive Neuroscience* [online] 23 (1), 183–199. https://doi.org/10.1162/jocn.2010.21425

Ojima, S., Nakata, H. and Kakigi, R. (2005). An ERP study of second language learning after childhood: effects of proficiency. *Journal of Cognitive Neuroscience* [online] 17 (8), 1212–1228. https://doi.org/10.1162/0898929055002436

Pagano, M. and Gauvreau, K. (2018). *Principles of biostatistics* (2nd edition). New York: Chapman and Hall/CRC.

Palomero-Gallagher, N., Amunts, K. and Zilles, K. (2015). Transmitter receptor distribution in the human brain. In A. W. Toga (ed.), *Brain mapping: an encyclopedic reference*. San Diego, CA: Academic Press, 261–275.

Patterson, K. and Lambon Ralph, M. A. (2016). The hub-and-spoke hypothesis of semantic memory. In G. Hickok and S. Small (eds.), *Neurobiology of language*. San Diego, CA: Academic Press, 765–775.

Pekár, S. and Brabec, M. (2016). Marginal models via GLS: a convenient yet neglected tool for the analysis of correlated data in the behavioural sciences. *Ethology* [online] 122 (8), 621–631. https://doi.org/10.1111/eth.12514

Pérez, A., Dumas, G., Karadag, M. and Duñabeitia, J. A. (2019). Differential brain-to-brain entrainment while speaking and listening in native and foreign languages. *Cortex* [online] 111, 303–315. https://doi.org/10.1016/j.cortex.2018.11.026

Phillips, N. A., Segalowitz, N., O'Brien, I. and Yamasaki, N. (2004). Semantic priming in a first and second language: evidence from reaction time variability and event-related brain potentials. *Journal of Neurolinguistics* [online] 17 (2-3), 237–262. https://doi.org/10.1016/S0911-6044(03)00055-1

Pu, H., Holcomb, P. J. and Midgley, K. J. (2016). Neural changes underlying early stages of L2 vocabulary acquisition. *Journal of Neurolinguistics* [online] 40, 55–65. https://doi.org/10.1016/j.jneuroling.2016.05.002

Purves, D., Augustine, G. J., Fitzpatrick, D., Hall. W. C., LaMantia, A. and White, L. E. (2012). *Neuroscience* (5th edition). Sunderland, MA: Sinauer Associates.

Pylkkänen, L. (2019). The neural basis of combinatory syntax and semantics. *Science* [online] 366 (6461), 62–66. https://doi.org/10.1126/science.aax0050

Pylkkänen, L. (2020). Neural basis of basic composition: what we have learned from the red–boat studies and their extensions. *Philosophical Transactions of the Royal Society B* [online] 375 (1791), n. pag. https://doi.org/10.1098/rstb.2019.0299

Rabovsky, M., Hansen, S. S. and McClelland, J. L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour* [online] 2 (9), 693–705. https://doi.org/10.1038/s41562-018-0406-4
Renoult, L., Irish, M., Moscovitch, M. and Rugg, M. D. (2019). From knowing to remembering: the semantic–episodic distinction. *Trends in Cognitive Sciences* [online] 23 (12), 1041–1057. https://doi.org/10.1016/j.tics.2019.09.008
Roig, M., Nordbrandt, S., Geertsen, S. S. and Nielsen, J. B. (2013). The effects of cardiovascular exercise on human memory: a review with meta-analysis. *Neuroscience & Biobehavioral Reviews* [online] 37 (8), 1645–1666. https://doi.org/10.1016/j.neubiorev.2013.06.012
Rubin, J. (1975). What the "good language learner" can teach us. *TESOL Quarterly* [online] 9 (1), 41–51. https://doi.org/10.2307/3586011
Rudland, J. R., Golding, C. and Wilkinson, T. J. (2020). The stress paradox: how stress can be good for learning. *Medical Education* [online] 54 (1), 40–45. https://doi.org/10.1111/medu.13830
Rudy, J. W. (2014). *The neurobiology of learning and memory* (2nd edition.). Sunderland, MA: Sinauer Associates.
Ruzich, E., Crespo-García, M., Dalal, S. S. and Schneiderman, J. F. (2019). Characterizing hippocampal dynamics with MEG: a systematic review and evidence-based guidelines. *Human Brain Mapping* [online] 40 (4), 1353–1375. https://doi.org/10.1002/hbm.24445
Sassenhagen, J. and Alday, P. M. (2016). A common misapplication of statistical inference: nuisance control with null-hypothesis significance tests. *Brain and Language* [online] 162, 42–45. https://doi.org/10.1016/j.bandl.2016.08.001
Schirmer, A., Soh, Y. H., Penney, T. B. and Wyse, L. (2011). Perceptual and conceptual priming of environmental sounds. *Journal of Cognitive Neuroscience* [online] 23 (11), 3241–3253. https://doi.org/10.1162/jocn.2011.21623
Sitnikova, T., Holcomb, P. J., Kiyonaga, K. A. and Kuperberg, G. R. (2008). Two neurocognitive mechanisms of semantic integration during the comprehension of visual real-world events. *Journal of Cognitive Neuroscience* [online] 20 (11), 2037–2057. https://doi.org/10.1162/jocn.2008.20143
Smith, M. E. and Farah, M. J. (2011). Are prescription stimulants "smart pills"? The epidemiology and cognitive neuroscience of prescription stimulant use by normal healthy individuals. *Psychological Bulletin* [online] 137 (5), 717–741. https://doi.org/10.1037/a0023825
Smith, S. M. and Nichols, T. E. (2009). Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage* [online] 44 (1), 83–98. https://doi.org/10.1016/j.neuroimage.2008.03.061
Smith, S. M. and Vale, W. W. (2006). The role of the hypothalamic-pituitary-adrenal axis in neuroendocrine responses to stress. *Dialogues in Clinical Neuroscience* [online] 8 (4), 383–395. https://doi.org/10.31887/DCNS.2006.8.4/ssmith
Soskey, L., Holcomb, P. J. and Midgley, K. J. (2016). Language effects in second-language learners: a longitudinal electrophysiological study of Spanish

classroom learning. *Brain Research* [online] 1646, 44–52. https://doi.org/10.1016/j.brainres.2016.05.028
Šoškić, A., Jovanović, V., Styles, S. J., Kappenman, E. S. and Ković, V. (2021). How to do better N400 studies: reproducibility, consistency and adherence to research standards in the existing literature. *Neuropsychology Review* [online] 1–24. https://doi.org/10.1007/s11065-021-09513-4
Squire, L. R. (2004). Memory systems of the brain: a brief history and current perspective. *Neurobiology of Learning and Memory* [online] 82 (3), 171–177. https://doi.org/10.1016/j.nlm.2004.06.005
Stelmack, R. M. and Miles, J. (1990). The effect of picture priming on event-related potentials of normal and disabled readers during a word recognition memory task. *Journal of Clinical and Experimental Neuropsychology* [online] 12 (6), 887–903. https://doi.org/10.1080/01688639008401029
Strimbu, K. and Tavel, J. A. (2010). What are biomarkers? *Current Opinion in HIV and AIDS* [online] 5 (6), 463–466. https://doi.org/10.1097/COH.0b013e32833ed177
Sweatt, J. D. (2016). Neural plasticity and behavior – sixty years of conceptual advances. *Journal of Neurochemistry* [online] 139, 179–199. https://doi.org/10.1111/jnc.13580
Teasdale, A. and Leung, C. (2000). Teacher assessment and psychometric theory: a case of paradigm crossing? *Language Testing* [online] 17 (2), 163–184. https://doi.org/10.1177/026553220001700204
Teige, C., Mollo, G., Millman, R., Savill, N., Smallwood, J., Cornelissen, P. L. and Jefferies, E. (2018). Dynamic semantic cognition: characterising coherent and controlled conceptual retrieval through time using magnetoencephalography and chronometric transcranial magnetic stimulation. *Cortex* [online] 103, 329–349. https://doi.org/10.1016/j.cortex.2018.03.024
Thomas, M. S., Ansari, D. and Knowland, V. C. (2019). Annual research review: educational neuroscience: progress and prospects. *Journal of Child Psychology and Psychiatry* [online] 60 (4), 477–492. https://doi.org/10.1111/jcpp.12973
Tononi, G. and Cirelli, C. (2014). Sleep and the price of plasticity: from synaptic and cellular homeostasis to memory consolidation and integration. *Neuron* [online] 81 (1), 12–34. https://doi.org/10.1016/j.neuron.2013.12.025
Traugott, E. C. (2017). Semantic change. In *Oxford research encyclopedia of linguistics*, n. pag.
Tulving, E. (1972). Episodic and semantic memory. In E. Tulving and W. Donaldson (eds.), *Organization of memory*. Oxford: Academic Press, 381–403.
Valbuena, S. and Lerma, J. (2021). Kainate receptors, homeostatic gatekeepers of synaptic plasticity. *Neuroscience* [online] 456, 17–26. https://doi.org/10.1016/j.neuroscience.2019.11.050
Van Lier, L. (2004). *The ecology and semiotics of language learning: a sociocultural perspective*. Boston: Springer.
Wacewicz, S. and Żywiczyński, P. (2015). Language evolution: why Hockett's design features are a non-starter. *Biosemiotics* [online] 8 (1), 29–46. https://doi.org/10.1007/s12304-014-9203-2

Wacewicz, S. and Żywiczyński, P. (2018). Language origins: fitness consequences, platform of trust, cooperation, and turn-taking. *Interaction Studies* [online] 19 (1-2), 167–182. https://doi.org/10.1075/is.17031.wac

Wang, L., Jensen, O., van den Brink, D., Weder, N., Schoffelen, J.-M., Magyari, L., Hagoort, P. and Bastiaansen, M. (2012). Beta oscillations relate to the N400m during language comprehension. *Human Brain Mapping* [online] 33 (12), 2898–2912. https://doi.org/10.1002/hbm.21410

West, B. T., Welch, K. B. and Galecki, A. T. (2014). *Linear mixed models: a practical guide using statistical software* (2nd edition). Boca Raton: Chapman & Hall/CRC.

Westerlund, M. and Pylkkänen, L. (2014). The role of the left anterior temporal lobe in semantic composition vs. semantic memory. *Neuropsychologia* [online] 57, 59–70. https://doi.org/10.1016/j.neuropsychologia.2014.03.001

Wittgenstein, L. (1953). *Philosophical investigations.* New York: Macmillan.

Wright, T. S. and Cervetti, G. N. (2017). A systematic review of the research on vocabulary instruction that impacts text comprehension. *Reading Research Quarterly* [online] 52 (2), 203–226. https://doi.org/10.1002/rrq.163

Xu, W., Kolozsvari, O. B., Oostenveld, R. and Hämäläinen, J. A. (2020). Rapid changes in brain activity during learning of grapheme-phoneme associations in adults. *NeuroImage* [online] 220, n. pag. https://doi.org/10.1016/j.neuroimage.2020.117058

Yule, G. (2017). *The study of language* (6th edition). Cambridge, United Kingdom: Cambridge University Press.

Yum, Y. N., Midgley, K. J., Holcomb, P. J. and Grainger, J. (2014). An ERP study on initial second language vocabulary learning. *Psychophysiology* [online] 51 (4), 364–373. https://doi.org/10.1111/psyp.12183

Zhang, L. and Pylkkänen, L. (2015). The interplay of composition and concept specificity in the left anterior temporal lobe: an MEG study. *NeuroImage* [online] 111, 228–240. https://doi.org/10.1016/j.neuroimage.2015.02.028