

University of Groningen

Discovering the brain stages of lexical decision

Berberyan, Hermine S.; van Rijn, Hedderik; Borst, Jelmer P.

Published in:
Brain and Cognition

DOI:
[10.1016/j.bandc.2021.105786](https://doi.org/10.1016/j.bandc.2021.105786)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2021

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Berberyan, H. S., van Rijn, H., & Borst, J. P. (2021). Discovering the brain stages of lexical decision: Behavioral effects originate from a single neural decision process. *Brain and Cognition*, 153, [105786]. <https://doi.org/10.1016/j.bandc.2021.105786>

Copyright

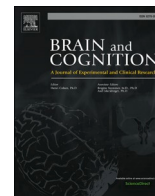
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.



Discovering the brain stages of lexical decision: Behavioral effects originate from a single neural decision process

Hermine S. Berberyan^{a,*}, Hedderik van Rijn^b, Jelmer P. Borst^a

^a Bernoulli Institute, University of Groningen, Groningen, the Netherlands

^b Department of Experimental Psychology, University of Groningen, Groningen, the Netherlands

ARTICLE INFO

Keywords:

Lexical decision
Electroencephalography
Event-related potentials
Machine learning
Multivariate pattern analysis
Hidden semi-Markov Models

ABSTRACT

Lexical decision (LD) – judging whether a sequence of letters constitutes a word – has been widely investigated. In a typical lexical decision task (LDT), participants are asked to respond whether a sequence of letters is an actual word or a nonword. Although behavioral differences between types of words/nonwords have been robustly detected in LDT, there is an ongoing discussion about the exact cognitive processes that underlie the word identification process in this task. To obtain data-driven evidence on the underlying processes, we recorded electroencephalographic (EEG) data and applied a novel machine-learning method, hidden semi-Markov model multivariate pattern analysis (HsMM-MVPA). In the current study, participants performed an LDT in which we varied the frequency of words (high, low frequency) and “wordlikeness” of non-words (pseudowords, random non-words). The results revealed that models with six processing stages accounted best for the data in all conditions. While most stages were shared, Stage 5 differed between conditions. Together, these results indicate that the differences in word frequency and lexicality effects are driven by a single cognitive processing stage. Based on its latency and topology, we interpret this stage as a *Decision* process during which participants discriminate between words and nonwords using activated lexical information.

1. Introduction

Language encompasses our daily life — whether it is a simple chat with family or friends, a more formal presentation at work, or reading the newspaper. One aspect involved in language processing is recognizing words from a sequence of phonemes or a string of letters. A common paradigm for studying word recognition is a lexical decision task (LDT; Meyer & Schvaneveldt, 1971). In this paradigm, participants are asked to indicate whether a combination of letters presented on the screen is an existing word or a nonword. A robust finding in LD is that the frequency-of-occurrence of words and “wordlikeness” of nonwords determine the speed and accuracy of responses (e.g., Adams, 1979; Balota & Chumbley, 1984; Brysbaert et al., 2011; Grainger, 1990; Perea, Rosa, & Gómez, 2005; Ratcliff, McKoon, & Gomez, 2004; Schilling, Rayner, & Chumbley, 1998). Yet, there is an ongoing discussion about the cognitive processes that cause these differences.

In this research, we aim to obtain evidence on the underlying cognitive processes of LDT. To that end, we will apply a novel machine-learning algorithm: hidden semi-Markov model multivariate pattern analysis (HsMM-MVPA; Anderson, Zhang, Borst, & Walsh, 2016). This is

a data-driven method that can discover processing stages from electroencephalographic (EEG) data. Using HsMM-MVPA, one can detect how many cognitive stages drive performance, and characterize these stages based on their latency and topology. Unlike two common methods previously used to analyze LDT – behavioral analyses that use a single measure per trial and event-related potentials (ERPs) that use averages of the EEG signal over trials and participants – HsMM-MVPA takes all data of the experiment into account simultaneously. As a result, HsMM-MVPA allows finding *the exact point in the trial* at which the differences between conditions occur and infer *which cognitive processes* they are associated with.

In the remainder of the introduction, we will present the main findings from the lexical-decision literature and describe the rationale for HsMM-MVPA analysis applied to EEG data.

1.1. Lexical decision task

One important phenomenon that has been observed in LD is the *word frequency effect* – words that occur more frequently in the lexicon (high frequency words; HF) are on average recognized more quickly than less

* Corresponding author.

E-mail address: h.berberyan@rug.nl (H.S. Berberyan).

frequent words (low frequency words; LF; e.g., Balota & Chumbley, 1984; Fiez, Balota, Raichle, & Petersen, 1999; Scarborough, Cortese, & Scarborough, 1977). Another robust phenomenon is the *nonword lexicality effect* that reflects that “wordlike” nonwords (pseudowords) are recognized more slowly than random nonwords (Fiez et al., 1999; Ratcliff et al., 2004; Stone & Van Orden, 1993; Wagenmakers et al., 2004).

Traditionally, the observed differences between HF and LF words were attributed to lexical access, that is, the process of retrieving the word from the mental lexicon (Becker, 1980; Donkin & Heathcote, 2009; Morton, 1969; Paap, Newsome, McDonald, & Schvaneveldt, 1982; Perea, Rosa, & Gómez, 2002; Schilling et al., 1998). In these models, the assumption was made that various word identification tasks share the same underlying process, lexical access, with a minor role for decision-related processes (Andrews & Heathcote, 2001). For example, in the conventional logogen model (Morton, 1969), the word frequency effect is due to a higher resting activation level of lexical units in HF words than in LF words (for an overview of models of lexical access, see Andrews, 1989). Paap et al. (1982) suggested a similar process for pseudowords, where words that are orthographically similar to the pseudowords are being retrieved from the lexicon. This explains why pseudowords are often mistaken with real words and are characterized by lower accuracy than nonwords (Grainger, Bouttevin, Truc, Bastien, & Ziegler, 2003).

In reaction to the finding that the word frequency effect was larger in the LDT than in other word recognition tasks, an alternative view was proposed (Balota & Chumbley, 1984; Besner, 1983; McCann, Besner, & Davelaar, 1988; Ratcliff et al., 2004; Stone & Van Orden, 1993; Wagenmakers, Ratcliff, Gomez, & McKoon, 2008). For example, Balota and Chumbley (1984) suggested that the differences in the LDT are driven by task-specific decisional or familiarity processes, rather than pure lexical access. In this frame, Ratcliff and colleagues (2004) explained the effects of word frequency and type of nonwords as altering the “amount and kind of information ... that drives the decision process and nothing more” (p. 20). Continuing this idea, Wagenmakers and colleagues (2008) showed that strategic factors also influence performance on the LDT by manipulating speed versus accuracy instructions.

Although mean reaction time (RTs) measures have been traditionally used to investigate frequency and lexicality effects, more sophisticated methodologies are needed to identify the processes underlying LD. Recent studies demonstrated that not only the average RT measure is indicative of these effects, but the shape of the whole RT distribution (Andrews & Heathcote, 2001; Balota & Spieler, 1999; Perea et al., 2005; Tillman, Osth, van Ravenzwaaij, & Heathcote, 2017). An alternative approach is the signal-to-respond paradigm where accuracy rather than speed of response was used as the variable of interest (Hintzman & Curran, 1997; Wagenmakers et al., 2004; Zeelenberg, Wagenmakers, & Shiffrin, 2004). However, even these more advanced behavioral methods could not clearly identify underlying cognitive processes. Thus, for example, Balota and Spieler (1999) found that the word frequency and lexicality had a different effect on RT distribution for naming and LD tasks. Although these differences were interpreted as the presence of the familiarity processes in LDT, there was no clear support for this.

With the advent of neuroimaging, researchers turned from behavioral measures to functional imaging. In this frame, functional Magnetic Resonance Imaging (fMRI) was used to identify the brain networks underlying the word recognition process in LDT. In a recent study on LDT and naturalistic reading, common reading-related regions including ventral occipito-temporal cortex and inferior frontal gyrus (IFG) were found (Wang et al., 2015). Additionally, task-specific activations in precentral gyrus and insula were detected in LDT, linked to motor preparation and error monitoring, respectively (Taylor, Stern, & Gehring, 2007). In another study on LDT and reading aloud, it was found that both tasks shared a wide variety of common brain areas associated with orthographic, semantic and phonological processing (Carreiras, Mechelli, Estévez, & Price, 2007). Moreover, in LDT, task-specific

activations were reported at bilateral postcentral gyri, supplementary motor area, and right cerebellum linked to decision-making and motor planning. Overall, this suggests that besides common reading processes, there are task-specific cognitive processes involved in LDT associated with decision making.

Due to its advantage of very high temporal resolution, EEG has been applied to LDT to investigate the temporal order of the different underlying processes. Because it is difficult to extract information about cognitive processing from raw EEG data, event-related potentials (ERPs), averages of the signal locked to a particular temporal event, were calculated from LDT-EEG data. One of the most robust findings was the N400 component, which negative amplitude was seen as the inverse of word frequency (Barber & Kutas, 2007; Barber, Vergara, & Carreiras, 2004; Kutas & Federmeier, 2011; Rugg, 1990). This component has also been observed for nonwords lexicality effects with pseudowords eliciting more negative amplitudes than words (Barber, Otten, Kousta, & Vigliocco, 2013; Carreiras, Vergara, & Barber, 2005; Holcomb, Grainger, & O'Rourke, 2002; Lau, Phillips, & Poeppel, 2008) and random non-words (Holcomb & Neville, 1990; Nobre & McCarthy, 1994).

Another ERP component that was found to vary between words/nonwords is a positively going P600, also referred to as the Late Positive Component (LPC). It has been found that words are characterized by a higher amplitude P600 as compared to pseudowords (Bermúdez-Margaretto, Beltrán, Domínguez, & Cuetos, 2015; Curran, 1999; Proverbio & Adorni, 2008). In studies where a task different than LDT was used, P600 was found to be associated with word frequency (Rugg, 1990; Young & Rugg, 1992). Another interesting finding was that not only the amplitude was associated with these effects but also the latency: the P600 was found to be longer for LF than HF words (Polich & Donchin, 1988).¹

Although amplitude effects of these components are well established, the temporal locations are less well defined. For example, for the negative-going N400, not only the reported interval is large (>200 ms) but also different between studies: 250–450 ms (Barber et al., 2013); 300–500 ms (Rugg, 1990); 350–500 ms (Barber et al., 2004); 250–500 ms (Nobre & McCarthy, 1994); 300–650 ms (Carreiras, Vergara, & Perea, 2007). For the P600, similar timing differences were observed: 400–700 ms (Curran, 1999); 500–618 ms (Bermúdez-Margaretto et al., 2015); 600–800 ms (Proverbio & Adorni, 2008); 500–800 ms (Young & Rugg, 1992). An important reason for this is the trial-by-trial variability in these peaks, which is more pronounced further from fixed time points (Borst & Anderson, in press; Walsh, Gunzelmann, & Anderson, 2017; Zhang, Walsh, & Anderson, 2018). The more variable a peak is between trials, the wider the component will be when calculating ERPs.

Overall, this makes it difficult to conclude at what particular moments in time differences between types of words and nonwords originate, and consequently, what the driving cognitive process is.

1.2. HsMM-MVPA analysis for discovering stages in EEG

Despite the fact that the traditional ERP approach has been instrumental in quantifying the differences in the word recognition process across words and nonwords, it has been limited in identifying the exact temporal locations where different stimuli elicit different brain signatures. If we could use the high temporal resolution of EEG data and precisely identify the timing of these peaks on each trial, then we could address *at which point in the trial the differences between various types of words and nonwords arise*. These differences could be further explored and associated with the underlying cognitive processes. Here, we will use a novel machine learning algorithm, HsMM-MVPA, that uses the

¹ Although labelled as P300 in this study, this positive-going component was characterized with a peak around 600 ms and most prominent positivity at centro-parietal regions consistent with findings on the P600.

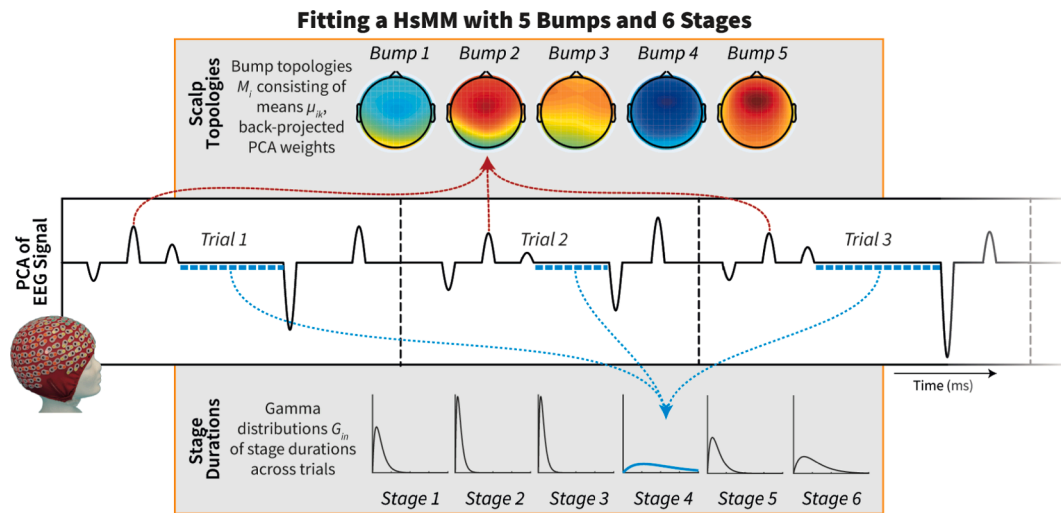


Fig. 1. A schematic representation of the HsMM-MVPA method applied to EEG data. This model consists of five bumps and six stages. Red lines indicate that the plotted topology for Bump 2 originates from the second peak present on each trial. Blue lines indicate a wide distribution of Stage 4 that originates from the differences in the duration of this stage across trials. Reprinted from Berberyan et al. (2021). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

EEG data to detect peaks on a trial-by-trial basis.

HsMM-MVPA is a method that allows parsing neural recordings of a cognitive task into different stages that are assumed to represent different, sequential processes (Anderson et al., 2016; Borst & Anderson, in press). To find these stages, Hidden semi-Markov models (HsMMs; Yu, 2010) were combined with multivariate pattern analysis. HsMMs are an extension of standard Hidden Markov Models (HMM; Rabiner, 1989) that allow for variable-duration stages. The onset of each stage is defined as a multivariate pattern across the EEG electrodes (when averaged, parts of these patterns form traditional ERPs). This HsMM-MVPA method has been successfully applied to a range of cognitive tasks including associative memory recognition, arithmetic retrieval, and classical short-term memory tasks (Anderson et al., 2018, 2016; Portoles, Borst, & van Vugt, 2018; van Maanen, Portoles, & Borst, 2021; Walsh et al., 2017; Zhang, Borst, Kass, & Anderson, 2017; Zhang, van Vugt, Borst, & Anderson, 2018; Zhang, Walsh, et al., 2018). In a recent validation study, it was found that HsMM-MVPA allows for zooming into processes identified with RT-based methods by finding a range of corresponding cognitive stages (Berberyan, van Maanen, van Rijn, & Borst, 2021).

The rationale for HsMM-MVPA analysis is the assumption that a cognitive event (i.e., the trigger for a new stage) is accompanied by a peak in the activity in different brain regions. This assumption is in line with two competing theories of ERP generation: *the classical theory* (Schroeder et al., 1995; Shah et al., 2004) and *the synchronized oscillations theory* (Makeig et al., 2002), which both assume a peak at the onset of a significant cognitive event. Consistent with this idea, we search for a multivariate peak in activation (*a bump*) to identify a new cognitive stage. Each bump marks the onset of a new processing stage and is followed by ongoing EEG oscillations that are referred to as *a flat*. There are multiple cognitive events in a task and, thus, multiple stages. A stage in HsMM-MVPA analysis is viewed as a combination of a bump and a flat, except for the first stage, which starts with stimulus onset and ends at the first bump.

In HsMM-MVPA analysis, an average model is created for all participants and trials. The underlying idea is that participants go through the same cognitive processes, and thus bumps, on each trial. However, as the duration of these processes can vary over participants and between trials, the flats between bumps have variable durations which are modelled as gamma distributions. This is illustrated in Fig. 1, where the red arrows indicate the origin of the second bump, and the blue arrows indicate the origin of the gamma distribution describing Stage 4. The method

first integrates the data of all participants and trials to define the optimal bumps and gamma distributions that account for the EEG data, and as a result, yields the most probable temporal location of each bump on each trial – allowing for a more detailed analysis of these bumps than is possible with traditional ERPs.

To summarize, in the current EEG study, we aimed to identify the cognitive stages that underlie the LD process and to determine which stages drive the difference between HF, LF, pseudowords, and non-words. To compare our results with the existing literature, we first performed standard behavioral and ERP analyses. Next, we applied the HsMM-MVPA analysis to discover the underlying processing stages. Finally, we calculated bump-related potentials (BRPs) – an extension of classical ERP analysis – by aligning EEG data to the temporal onsets of these stages (Borst & Anderson, in press).

2. Methods

2.1. Participants

Twenty-nine volunteers, students of the University of Groningen, performed the LDT. Three participants were excluded from the analysis: one due to recording problems and two – due to excessive noise in the EEG data. The final set consisted of twenty-six participants (11 females, age range 18–35 years old, mean age = 22.5, SD = 3.2). For their participation, they received a compensation of 8 euros.

All participants were right-handed, had no history of neurological disorders and were Dutch native speakers. Before the experiment, all participants gave written informed consent in accordance with the Declaration of Helsinki. The experimental procedures were approved by the Research Ethics Review Committee (CETO) of the Faculty of Arts at the University of Groningen (reference number 64329605).

2.2. Task

In the experiment, participants were asked to respond whether a combination of letters presented on the screen was an actual Dutch word or not. To respond, participants pressed the corresponding button on the keyboard ('n' or 'm') with their right hand. If the response was incorrect they received feedback. We varied the natural frequency of the words (HF and LF) and the 'wordlikeness' of non-words (pseudowords and random letter strings).

2.3. Materials

To ensure that our results reflect standard LD effects, we based our stimuli on a previous dataset. We used items from the Dutch Lexicon Project (DFP; Keuleers, Diependaele, & Brysbaert, 2010) that included 14,000 words and 14,000 nonwords. From these, we first selected items with 4–8 characters that had at least 85% accuracy. For the HF condition, we then selected items with a CELEX² frequency (Baayen, Piepenbrock, & van Rijn, 1993) higher than 1000 (mean = 7352.10, SD = 22693.64, $n = 1702$). For the LF condition, we selected words with a CELEX frequency higher than 3 and lower than 25 (mean = 12.81, SD = 6.17, $n = 506$). For the Pseudoword condition, we selected nonwords that had a corresponding HF item with a maximum edit distance of 2 and consisted of the same number of characters. For the Random letter strings condition, we created new items by randomly shuffling characters of the LF items and making sure that the new strings did not accidentally form a word. Finally, for each participant, we randomly selected 240 items for each condition and 8 practice items.

2.4. Procedure

Participants were seated in front of a 21.5-inch screen. The experiment was implemented in OpenSesame (Mathôt, Schreij, & Theeuwes, 2012) using the *xpyriment* back-end. The experiment consisted of 960 trials presented in four blocks. Each of these blocks contained 240 trials, 60 trials per condition. A practice session of 8 trials preceded the main experimental session. The main experimental part took on average 30–40 min. The entire duration of the experiment including instructions and EEG setup was one hour.

Each trial started with a fixation dot for a random duration between 400 and 600 ms. Next, the stimulus (word/non-word) was presented in 20-points font size. The participants were asked to respond by pressing the button on the keyboard - 'n' for words and 'm' for non-words. If participants responded incorrectly or did not respond during a specified timeout (2000 ms), the feedback 'Incorrect' was presented for 2000 ms.

2.5. EEG recording

EEG was acquired from 32 positions using active Ag-AgCl electrodes (Biosemi Active Two system). Data was recorded with a sampling rate of 512 Hz. The electrodes were placed using the international 10–20 system layout. Two channels – Common Mode Sense (CMS) and Driven Right Leg (DRL) – were used as “ground”. Four additional electrodes (two vertical and two horizontal) were used to measure eye movements. For most of the participants, scalp impedance of the electrodes was kept at <20 k Ω , while for six participants it was kept at <30 k Ω .

2.6. Behavioral analysis

Practice trials were excluded for both accuracy and reaction times (RT) analyses. Additionally, for the RT analysis, we excluded all incorrect responses. For each condition and each participant, we then removed trials that deviated more than 3 standard deviations from the mean RTs (1.76% of the trials). This form of outlier rejection was in line with previous HsMM-MVPA studies (e.g., Anderson et al., 2016).

Differences in accuracy and RTs were evaluated with linear mixed-effects models (LMEs; (Bates & DebRoy, 2004) as implemented in R. LMEs are extensions of linear regression models that allow to flexibly combine both fixed and random effects. We obtained p -values for fixed effects, using Satterthwaite's method (Kuznetsova, Brockhoff, & Christensen, 2017). To obtain the random effects structure that could account best for the data, we fitted various models with random effects and compared the goodness-of-fit by means of likelihood-ratio tests. In this

way, we identified the maximum random-effects structure that was supported by the data (Bates, Kliegl, Vasishth, & Baayen, 2015).

2.7. EEG preprocessing

EEG preprocessing was conducted with the open source EEGLAB toolbox (Delorme & Makeig, 2004) and custom-made scripts in MATLAB. Data was first referenced to the average of the mastoids. Next, a high-pass filter of 0.5 Hz and a low-pass filter of 40 Hz were applied. To detect artifacts in the data manual artifact rejection was performed. On average 5.35 % of data was removed during manual artifact rejection. For ten participants, on average 1.5 noisy channels were detected and removed. Next, the data was decomposed with independent component analysis (ICA) using the infomax algorithm runica in EEGLAB. For all participants, one or two components containing eye blinks and muscle movements were identified and subtracted. Finally, the previously removed channels were interpolated with spherical spline interpolation.

2.8. ERP analysis

For ERP analysis, the ‘outliers’ rejection was consistent with the behavioral analysis pipeline (for details, see the 2.6. Behavioral analysis section). In addition, incomplete trials that appeared due to visual artifact rejection in the selected time windows were inspected and removed. Baseline normalization (400 ms preceding the stimulus presentation) was applied to remove slow drifts in the signal. The data from the following time intervals was selected: 200 ms before the stimulus and 600 ms after, 600 ms before the response and 200 ms after, for stimulus-locked and response-locked ERPs, respectively. Grand average event-related potential waveforms elicited by the four conditions were computed for all channels and visually inspected.

To statistically compare the observed differences between conditions, a cluster-based random permutation analysis was performed (Maris & Oostenveld, 2007) in Fieldtrip (Oostenveld, Fries, Maris, & Schoffelen, 2011). This analysis allows dealing with multiple comparison problems (MCP) and, thus, controlling for type 1 error (“false positive”). The general procedure is as follows: (1) for every sample, the difference between two conditions is calculated using a t -test, (2) clusters are formed based on the neighborhood, and the cluster statistic is calculated as the sum of the corresponding t -values, (3) the data is randomly divided into two conditions many times, and cluster statistics are calculated from that, (4) the proportion of the values from random partitions that exceeded the observed statistic is computed, resulting in a p -value (Monte Carlo approximation).

Because of the within-subject design, the data between conditions was compared by means of a paired t -test. Next, a cluster-level statistic was calculated as the sum of temporally-adjacent t -values that exceeded a p -threshold of 0.05. The data was randomly partitioned into two conditions 1000 times, and the histogram of cluster-level statistics was constructed. Next, the resulting p -value was calculated. The critical alpha level was specified as 0.05 meaning that if the p -value was smaller than 0.05, we concluded that the compared conditions were significantly different. The analysis was performed for all channels and conditions.

2.9. Preprocessing for HsMM-MVPA analysis

The trial selection for the HsMM-MVPA analysis was identical to the ERP analysis. For this analysis, the data from each individual trial was selected from stimulus presentation to response execution. Next, preprocessing specific to HsMM-MVPA analysis was performed (Anderson et al., 2016).

First, the data was downsampled to 100 Hz for computational efficiency, followed by a principal component analysis (PCA). First, the variance of each participant was scaled to the mean variance in the data to ensure that the data from all participants contributed equally to the

² Raw frequency that is specified by CELEX database

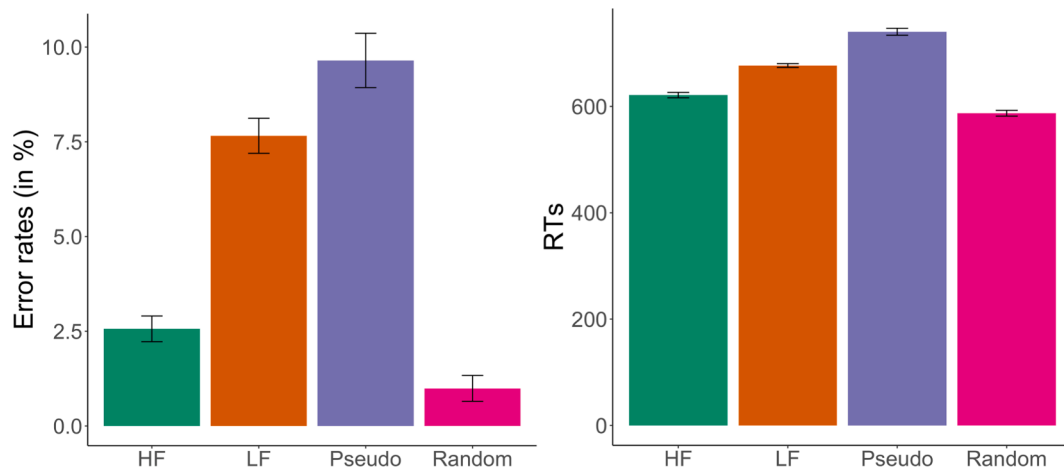


Fig. 2. Bar plots of error rates and reaction times are presented for each condition. The mean values are accompanied by within-subject standard errors (black lines).

results. Next, the covariance matrix was computed for each trial, and PCA analysis was performed on the average covariance matrices (Cohen, 2019; Portoles et al., 2018). Following previous HsMM-MVPA analyses, 10 PC components, which explained 90.5 % variance of the data, were used in the analysis. Finally, these PC components were z-scored per trial.

2.10. HsMM-MVPA application to EEG

The HsMM-MVPA analysis was applied to the EEG data to discover the processing stages underlying task performance. As described above (1.2. *HsMM-MVPA analysis for discovering stages in EEG*) a cognitive stage consists of a “bump” that is followed by a “flat”. Because bumps characterize cognitive processes that are assumed to be the same across all subjects, they have fixed topologies. On the other hand, the duration of these processes varies over trials, which is why the flats are described with gamma distributions.

In HsMM-MVPA analysis, several preliminary assumptions are made regarding bumps and flats. The width of the bumps is defined as 50 ms. Although the ‘real’ bumps could be narrower or wider than that, analyses on synthetic data revealed that the duration of 50 ms allows for robustly detecting bumps with a width of 30–110 ms (Anderson et al., 2016). The flats follow gamma distributions with a shape parameter of 2. While the shape parameter is fixed, the scale parameter of the gamma distribution is estimated. Bumps mark the transition to a new stage and thus initiate all stages except for the first stage that begins with stimulus presentation. Therefore, the placement of n bumps in a trial results in $n + 1$ flats.

HsMM-MVPA analysis is performed on all data simultaneously, and yields estimated bump locations per trial. The HsMM-MVPA analysis searches for the locations of the bumps, which are defined by a multivariate amplitude pattern. For maximum likelihood estimation, the standard expectation–maximization (EM) algorithm was used which is especially powerful when dealing with latent variables (Zhang, Brady, & Smith, 2001). As the expectation of EM, “E-Step”, we computed estimates for bumps (i.e., the amplitude) and flats (i.e., the scale parameter of the gamma distribution). As the maximization of EM, “M-Step”, we computed the maximum likelihood for these estimates. The steps repeat until the algorithm converges. This procedure results in estimates for the amplitude of the bumps, the scale parameters of the gamma distributions, and the probability of the location of each bump on each trial.

Our analysis started with the identification of good initial estimates for the parameters describing the bumps and gamma distributions, to avoid local maxima during subsequent EM estimations (Berbery et al., 2021; Zhang, Walsh, et al., 2018). This was done by first fitting an HsMM model with the maximum number of bumps to the data (as bumps

cannot overlap, this is the duration of the shortest trial divided by the 50-ms bump width). Next, models were fitted where one of these previously identified bumps was iteratively excluded. From these iterative models, we selected the one with the highest goodness-of-fit (i.e., log-likelihood). The described steps were repeated up to the model with 1 bump. The bump amplitudes and gamma distributions obtained in this way were then used in subsequent HsMM models as the starting parameters.

Next, to obtain the number of bumps that can account best for the data, HsMM models were fitted starting from 1 bump to the maximum number of bumps that fit in the data. Because the log-likelihood of a model typically increases when more bumps are fitted due to the increased flexibility of the model, a leave-one-out cross-validation (LOOCV) procedure was applied. For this aim, we fitted a model to all participants but one and tested the resulting parameters on this participant. The described procedure was repeated for all participants. A more complex model was preferred only if it outperformed a more parsimonious model for a significant number of participants.

2.11. BRP analysis

To extend the classical ERP analysis, we calculated bump-related potentials (BRPs). BRPs were calculated for each problem type by aligning EEG data to the most probable location of a bump on each trial identified with HsMM-MVPA. Next, for averaging purposes, the data was resampled to the average location of the bumps across trials. Finally, the grand average bump-related potential waveforms were calculated per condition. The differences between conditions were subjected to the same cluster-based random permutation analysis as the ERPs (for details, see 2.8. *ERP analysis* section).

3. Results

3.1. Behavioral results

Average error rates and RTs were computed for each participant and each condition. Next, average values per condition along with within-subject standard errors were calculated (Morey, 2008); these values are shown in Fig. 2. Reaction times were shortest for random non-words, followed by HF words, LF words, and pseudowords. Error rates followed the same pattern. These findings are consistent with the literature (e.g., Fiez et al., 1999; Ratcliff et al., 2004; Schilling et al., 1998; Wagenmakers et al., 2008) and the values previously reported for these stimuli (Keuleers et al., 2010).

To evaluate the observed differences statistically, LMEs were constructed. For both RTs and accuracy, a forward fitting procedure was

Table 1
The results of LME models for reaction times and accuracy.

RT ~ condition + (1 subject) + (1 item) + (0 + trial subject)					accuracy ~ condition + (1 subject) + (1 item) + (0 + trial subject)			
	B	SE	t	p	B	SE	z	p
Reference level: HF								
Intercept	623.04	12.80	48.66	<0.001	4.53	0.12	36.87	<0.001
LF	57.98	3.60	16.09	<0.001	-1.26	0.13	-9.57	<0.001
Pseudo	120.16	3.35	35.83	<0.001	-1.58	0.12	-13.47	<0.001
Random	-32.72	3.55	-9.21	<0.001	1.04	0.18	5.74	<0.001
Reference level: LF								
Intercept	681.02	12.90	52.78	<0.001	3.27	0.11	30.45	<0.001
Pseudo	62.18	3.72	16.73	<0.001	-0.31	0.11	-2.82	<0.01
Random	-90.69	3.89	-23.27	<0.001	2.30	0.18	12.72	<0.001
Reference level: Pseudo								
Intercept	743.20	12.84	57.90	<0.001	2.95	0.09	31.87	<0.001
Random	-152.87	3.67	-41.66	<0.001	2.62	0.17	15.36	<0.001

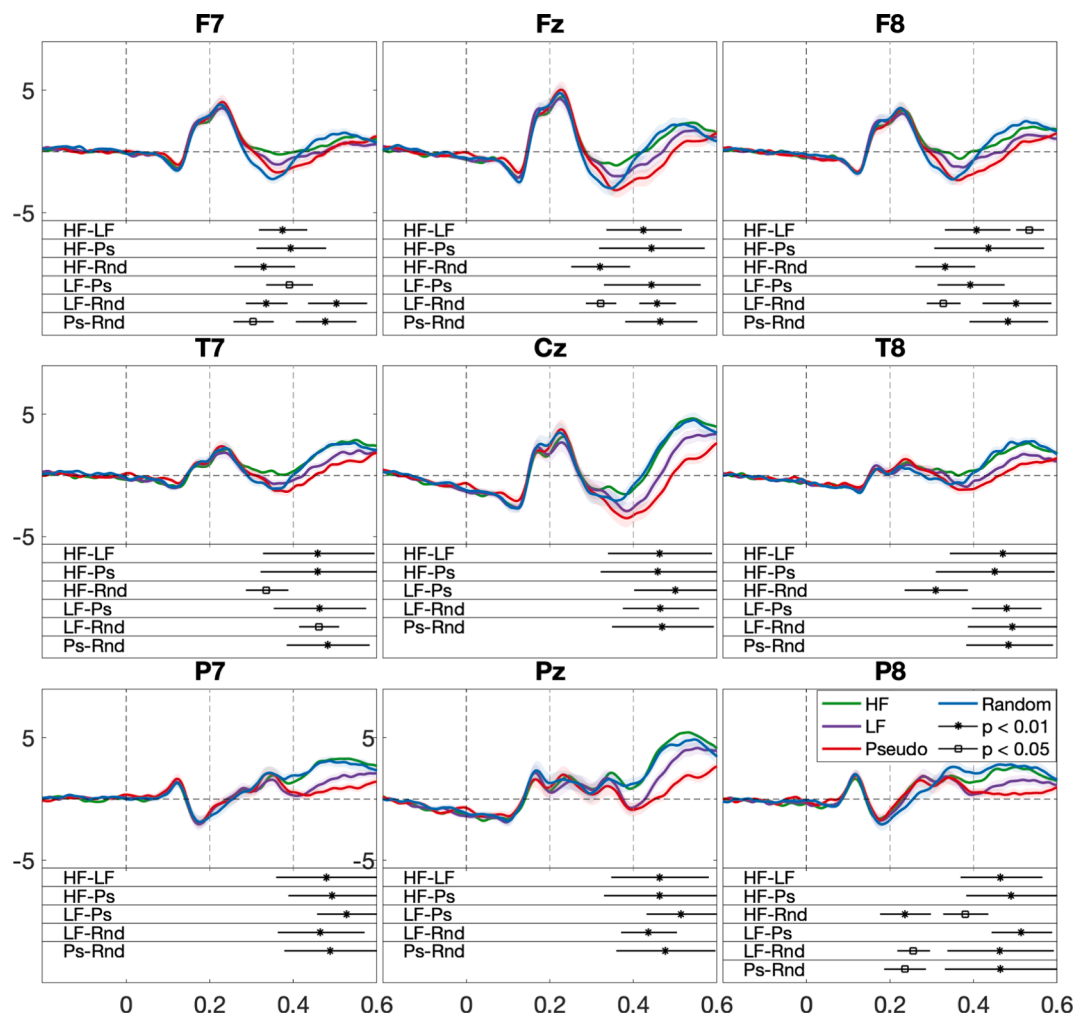


Fig. 3. Grand average ERPs for four conditions: HF (high frequency), LF (low frequency), Ps (pseudowords) and Rnd (random non-words) complemented with standard errors (shaded areas). The lines below the grand averages indicate significant clusters of differences between conditions. The location of these lines on the horizontal axis corresponds to the temporal locations of the significant differences. The clusters significant at 0.05 are marked with an asterisk (*) and at 0.01 with a rectangle (□).

applied. This revealed that the models with Condition as a fixed effect, Subject and Item as random intercepts and Trial as a random slope explained the most variance. For both accuracy and reaction times, significant differences were found between all conditions. The results are presented in Table 1.

3.2. ERP results

Grand average stimulus-locked ERPs were calculated for each condition. Fig. 3 shows 9 channels across the scalp, stimulus-locked (for response-locked ERPs, see Appendix 1). Significant clusters of condition differences are indicated with horizontal lines. The two main ERP components that are typically reported in LDT, N400 and P600, were

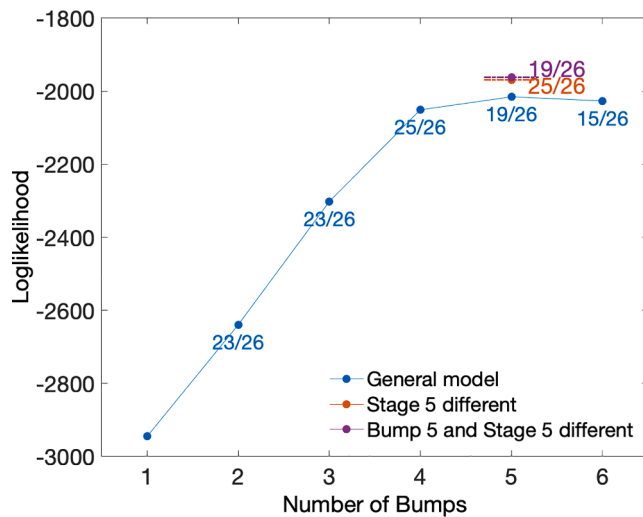


Fig. 4. The fits of the three models: the general model, a model where the duration of Stage 5 is assumed to differ between conditions, and a model where Bump 5 and the duration of Stage 5 are assumed to differ between conditions. The blue numbers associated with the general model indicate the number of participants for which the model improved as compared to a simpler model with fewer bumps (e.g. model with 2 bumps vs. a model with 1 bump). The values for “Stage 5 different” and “Bump 5 and Stage 5 different” models represent an improved fit as compared to simpler models with the same number of bumps. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2

The results of the LOOCV analysis. The rows in the table represent the type of model (general model or condition-based model) while columns represent the number of bumps in the current model as compared to a model with 1 bump less (e.g. $2 > 1$ means a 2-bump model compared to a 1-bump model). The values indicate the number of participants for which the current model outperformed the simpler model. Grey background indicates a significant increase (with at least 19 out of 26 subjects improving).

Type of model	Number of bumps				
	2 > 1	3 > 2	4 > 3	5 > 4	6 > 5
General model	23	23	25	19	15
HF model	25	24	25	23	14
LF model	25	21	25	24	10
Random model	26	23	25	25	18
Pseudo model	24	25	26	20	8

found.

Over central, parietal and temporal regions, the N400 was present with LF words eliciting more negative amplitudes than HF words and pseudowords being more negative than words. Our findings are consistent with the body of literature on the N400 linked to word frequency (e.g., Barber et al., 2004; Rugg, 1990). Moreover, the higher amplitude of the N400 was consistently reported for pseudowords as compared to words (e.g., Carreiras et al., 2005; Wang & Yuan, 2008) and to random nonwords (e.g., Coch & Mitra, 2010; Ziegler, Besson, Jacobs, Nazir, & Carr, 1997).

In addition, the positive-going P600 was observed. Most prominent at central and parietal locations, pseudowords elicited lower amplitude in the P600 as compared to random nonwords (Holcomb & Neville, 1990) and words (Ziegler et al., 1997). Additionally, HF words elicited more positive amplitudes in the P600 than LF words. This is in line with findings that the P600 is commonly associated with lexicality (e.g., Bermúdez-Margaretto et al., 2015; Holcomb & Neville, 1990) and, in some cases, with frequency effects (Young & Rugg, 1992).

3.3. HsMM-MVPA results

3.3.1. Number of stages in lexical decision

Under the assumption that cognitive stages are shared between the four different conditions: HF, LF, Random and Pseudowords, we started with a general model for all conditions. The results of the leave-one-out-cross-validation are shown as the blue line in Fig. 4. Five bumps – and thus six stages – were found to account best for the data, both in overall log-likelihood and in displaying an improvement over a solution with four bumps for a significant number of participants (19 of 26 participants, $p = 0.03$ as indicated by a sign test). In addition – to check whether this result was not due to pooling the data across conditions – we tested separate models for each condition. In all these models, five bumps were found to account best for the data (Table 2). We conclude that a model with five bumps and six stages best describes lexical decision, independent of condition.

3.3.2. Model selection

Initial visualisation of the models (not shown) revealed apparent differences in the duration of Stage 5 between conditions (the flexibility of the gamma distribution allows for this). To test whether the duration of this stage indeed needs to differ between conditions, we constructed two additional models: a model with only the duration of Stage 5 differing between conditions (i.e., each condition has its own gamma distribution for Stage 5) and a model with both Bump 5 and the duration of Stage 5 being different. The results are illustrated in Fig. 4 as the orange-red and the purple dots, respectively. The model where we varied the duration of Stage 5 across conditions clearly provided a better fit for a significant number of participants compared to the general model (25 of 26 participants, $p < 0.001$ as indicated by a sign test). The model where we varied both the duration of Stage 5 and Bump 5 had the highest log-likelihood (Fig. 4). Moreover, for a significant number of participants (19 of 26, $p = 0.03$), it outperformed the model where we varied only the duration of Stage 5.

For exploratory purposes, we also constructed models where the durations of other stages and the bumps varied per condition. None of these models outperformed the models above. On this basis, we conclude that a model with Bump 5 and Stage 5 being different across conditions while the other stages are shared, accounts best for the lexical decision process.

3.3.3. The resulting HsMM-MVPA model

The resulting HsMM-MVPA model is illustrated in Fig. 5. While Stages 1–4 and 6 are shared across conditions, Stage 5 is characterized by different durations between conditions, with the overall RT effect clearly originating in this stage. To evaluate these differences statistically, we constructed LMEs in R with the lmerTest package (Kuznetsova et al., 2017). A forward fitting procedure identified that the model with Condition as a fixed effect, Subject as a random intercept and Trial as a random slope explained the most variance. The duration of Stage 5 was different between all conditions (see Table 3 for summary statistics).

In the resulting HsMM-MVPA model, along with the fifth stage, the last bump varied between conditions and was characterized by a prominent parietal positivity. We compared the amplitude of Bump 5 for each channel and condition pair using paired t-tests followed by false discovery rate control across channels (FDR; Benjamini & Hochberg, 1995). Paired t-tests revealed significant differences between the following conditions: Pseudo-HF, Pseudo-LF, Random-HF and Random-Pseudo (see Fig. 6). While the differences between Random-HF and Random-Pseudo were observed only in very few channels, the differences between Pseudo-HF and Pseudo-LF were clustered in the spatial domain, mainly in the left parietal and occipital channels.

3.4. BRP results

Grand average BRPs were calculated for each condition and

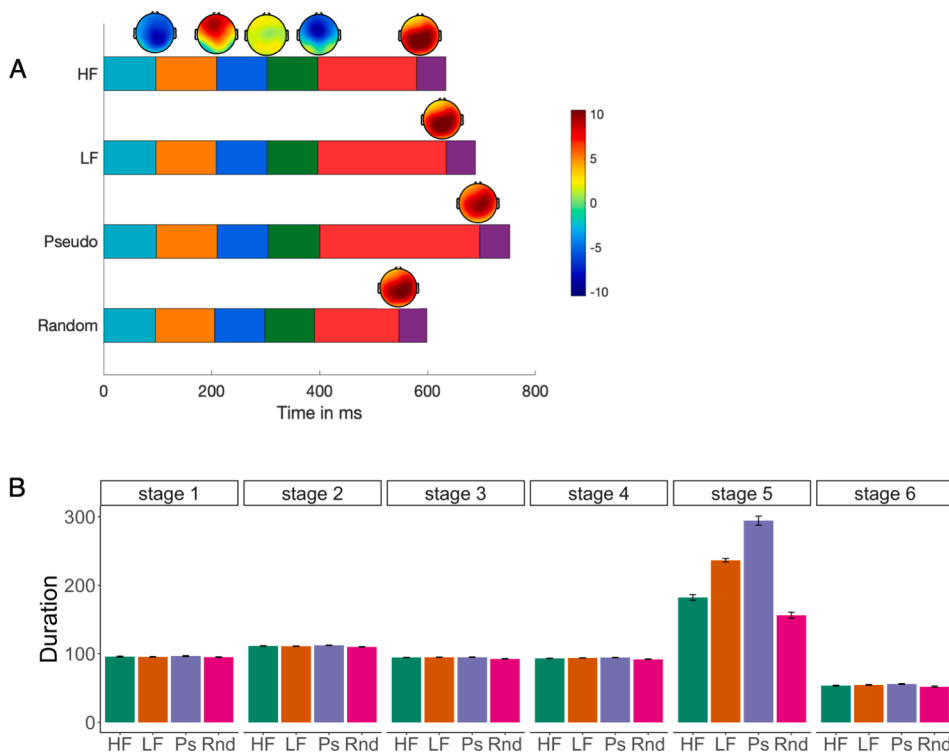


Fig. 5. The resulting HsMM-MVPA model where the duration of Stage 5 and Bump 5 differ across conditions. **A.** The first four topologies represent the amplitude of the bumps that are shared across all conditions, while the fifth bump is plotted separately for each condition. The bars below the topologies represent the average stage durations across participants and trials. **B.** The stage durations accompanied by within-subject standard errors (black lines) for each condition: HF (high frequency), LF (low frequency), Ps (pseudo-words) and Rnd (random non-words).

Table 3
The results of LME models for the duration of Stage 5.

Stage 5 ~ condition + (1 subject) + (0 + trial subject)				
	B	SE	t	p
<i>Reference level: HF</i>				
Intercept	183.23	8.27	22.16	<0.001
LF	53.46	2.30	23.18	<0.001
Pseudo	111.96	2.32	48.27	<0.001
Random	-26.26	2.27	-11.58	<0.001
<i>Reference level: LF</i>				
Intercept	236.69	8.27	28.60	<0.001
Pseudo	58.50	2.35	24.92	<0.001
Random	-79.71	2.29	-34.68	<0.001
<i>Reference level: Pseudo</i>				
Intercept	295.19	8.28	35.65	<0.001
Random	-138.22	2.31	-59.76	<0.001

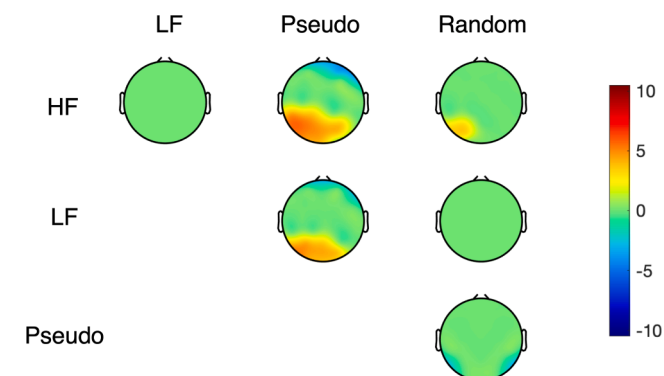


Fig. 6. A comparison of the amplitude of Bump 5 across conditions (row > column, e.g. HF > LF). The displayed topologies represent *t*-values for the significant differences that were corrected for false discovery rate (FDR; $p < 0.01$). The values for non-significant differences ($p > 0.01$) were set to 0.

presented for 9 channels across the scalp in Fig. 7. The cluster-based permutation analysis identified clusters of differences that are indicated with horizontal lines below the grand averages. Consistent with the ERP results, most of the differences in the amplitudes were observed in two consecutive ERP components: the N400 and P600. The N400 effect was observed for all conditions with pseudowords eliciting more negative amplitudes than words and random non-words. This component also accounted for frequency effects with more negative amplitudes for LF than HF. As in the ERPs, across central and parietal channels, pseudowords elicited lower amplitudes for P600 than the other conditions. In addition, the P600 component also accounted for frequency effects in some channels, particularly Pz.

The largest difference between ERPs and BRPs is that the latter allows for identifying the exact timings of the peaks for each condition. Most notably, not only the amplitude of the P600 peak was different between conditions but also its temporal location. This was not clear in the ERP results reported above, nor in the response-locked ERPs in the Appendix 1. These differences in the location of the P600 can directly be related to the HsMM-MVPA results: Stage 5 starts at the N400 and ends at the P600, and is clearly the underlying factor in the RT differences between the conditions.

4. Discussion

In the current study, we investigated the brain stages determining performance in lexical decision. While the LDT task has been widely investigated and has yielded robust behavioral and ERP findings, there has been an ongoing debate on the cognitive interpretation of these findings, in particular about the origin of the word frequency and nonword lexicality effects. To pinpoint the origin of these effects, we used a machine learning method that discovers the underlying cognitive stages directly from EEG.

To ensure that our results were comparable to standard LD findings, we based our stimuli on a previous dataset (Keuleers et al., 2010). Next, we performed a traditional behavioral analysis and calculated ERPs. Behavioral findings – RTs and error rates – displayed differences between all conditions in line with previous findings (Fiez et al., 1999;

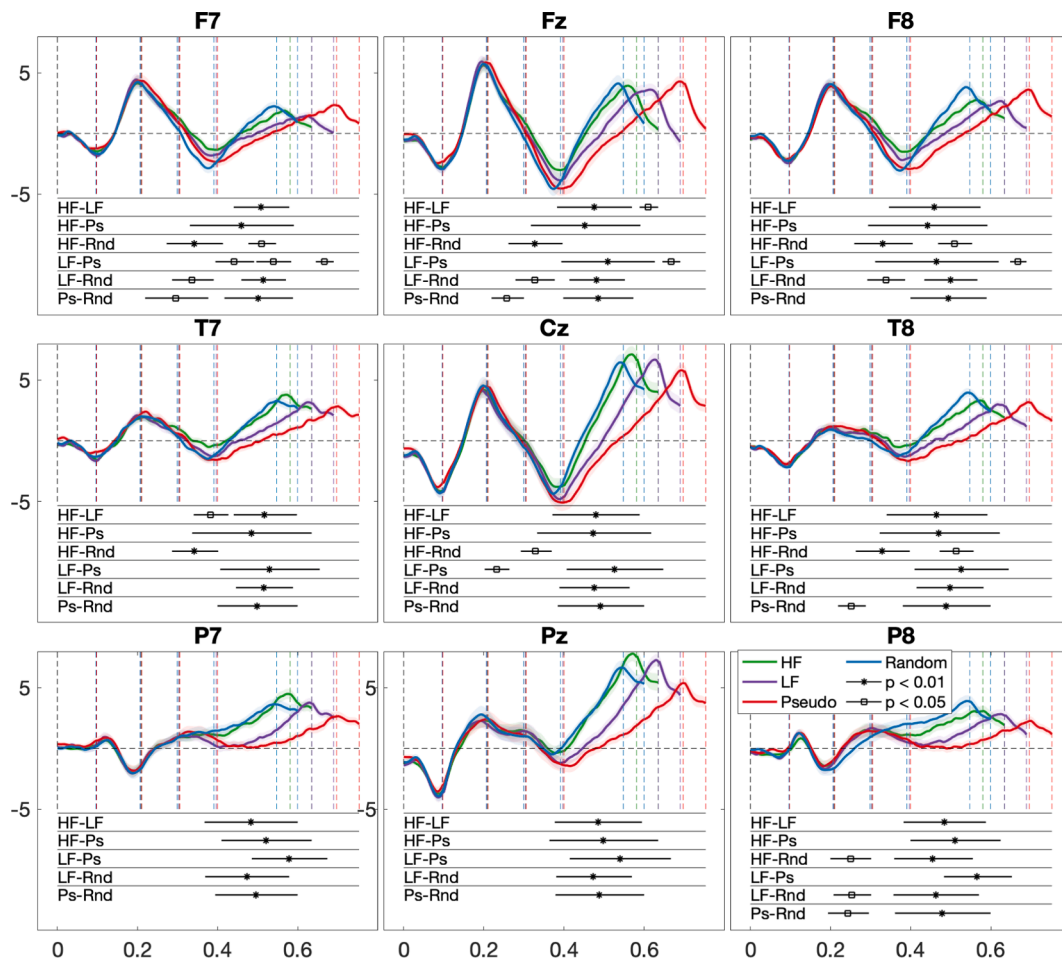


Fig. 7. Grand average BRPs for four conditions: HF (high frequency), LF (low frequency), Ps (pseudowords) and Rnd (random non-words) complemented with standard errors (shaded areas). The vertical lines indicate the average locations of the bumps for each condition marked with corresponding colours. The horizontal lines below grand averages indicate the significant clusters of differences between conditions. The clusters significant at 0.05 are marked with asterisk (*) and at 0.01 with a rectangle (□).

Schilling et al., 1998; Wagenmakers et al., 2008). Consistent with the literature, ERP results indicated two main components associated with these differences in RTs, the N400 and P600 (Carreiras et al., 2005; Holcomb & Neville, 1990).

To identify cognitive stages, the HsMM-MVPA analysis was applied to the EEG data, providing evidence for 5 bumps and 6 stages for all four conditions. While most of these stages were shared, Stage 5 and the consecutive Bump 5 were found to differ between conditions. The duration of Stage 5 was different for all conditions, while the amplitude of Bump 5 was greater for HF and LF items than for pseudowords. This leads to our first main conclusion: both the word frequency effect and the nonword lexicality effect originate in a single stage.

Based on the brain activations and durations, the discovered stages were further interpreted concerning the underlying cognitive processes. Stages 1–2 highly resemble the stages from other HsMM-MVPA studies where they were interpreted as pre-attention and early encoding stages (Anderson et al., 2016; Zhang, Walsh, & Anderson, 2017). This is in line with the findings on N100 and P200 components (Braun et al., 2006; Holcomb et al., 2002) and reading-related activations in occipito-temporal cortex and IFG (Murphy, Jorgia, & Talcott, 2019; Wang et al., 2015) previously reported in LDT.

The topological distribution and the duration of Stage 3 mirror a P300 (P300a) component typically associated with the allocation of attentional resources elicited by the stimuli (Polich, 2007; Soltani & Knight, 2000). Evidence suggests that the amplitude of this component is inversely related to the amount of resources required for the task (Kok,

2001). While neural origins of P300a are still not well defined, it was previously found that frontal lobe and hippocampus might play a role in its generation (Polich, 2007).

Stage 4 is identical to a familiarity-driven recognition stage previously reported in Berberyan et al. (2021). This pattern is consistent with familiarity-based theories where the familiarity of the stimulus was assumed to influence a consecutive lexical decision (Balota & Spieler, 1999; Wagenmakers et al., 2004; Zeelenberg et al., 2004). However, even though the familiarity of the words was probably judged in this stage, this did not affect the duration of the stage, in agreement with familiarity mechanisms in other tasks (e.g., Borst, Ghuman, & Anderson, 2016). The brain activation of this stage is compatible with recent ERP studies where familiarity was associated with frontal negativity (Curran, 1999; Woodruff, Hayama, & Rugg, 2006). fMRI studies have associated this activity with activation in prefrontal cortex (Aly, Yonelinas, Kishiyama, & Knight, 2011; Horn et al., 2016). However, a prominent alternative view argues that familiarity is related to the medial temporal lobe (e.g., Borst et al., 2016; Eichenbaum, Yonelinas, & Ranganath, 2007; Norman & O'Reilly, 2003).

The next question concerns the function of Stage 5. First, Stage 5 is the penultimate stage, which implies that it does not include motor preparation, as that is most likely located in the final stage. Second, in a previous study in which we experimentally manipulated decision difficulty, we saw a very similar bump-stage-bump pattern (Berberyan et al., 2021). In that study, the onset bump of the decision stage showed strong frontal negativity, and the end of the decision stage strong central-

parietal positivity – highly resemblant of the current bump topologies of bumps 4 and 5. When translating these findings to the current task, it implies that Stage 5 is a decision stage, where the difficulty of the decision determines the duration of the stage, as well as the amplitude of Bump 5.

The current patterns are in line with the conventional P300b effect, which is known to vary with response confidence, with higher confidence leading to more positive P300 responses (Borst & Anderson, 2015; Nieuwenhuis, Aston-Jones, & Cohen, 2005; Sutton, Ruchkin, Munson, Kietzman, & Hammer, 1982). This suggests that the decision was most difficult for pseudowords, which is corroborated by the RT and error rate results. On the level of the neural structures, these results are compatible with previously discussed fMRI results where bilateral postcentral gyri, supplementary motor area, and right cerebellum were reported in LDT representing task-specific decision and motor planning (Carreiras, Mechelli, et al., 2007).

Taken together, our findings are consistent with the previously discussed view that performance differences in the LDT can be best explained by differences in decision making (e.g., Balota & Chumbley, 1984; Besner, 1983; McCann et al., 1988; Ratcliff et al., 2004; Seidenberg & McClelland, 1989; Wagenmakers et al., 2004). Balota and Chumbley (1984) represented this decision process by a two-stage model where the first stage corresponded to a quick initial global computation of familiarity and the second - more analytical evaluation leading to decision. This “familiarity” process can be viewed as a result of various sources of lexical information such as orthographic, phonological and semantic information (Plaut, 1997). Continuing this idea, Ratcliff et al. (2004) suggested that these sources of information produce differences in a single component of drift diffusion model (DDM): drift rate, which has been typically associated with a decision (e.g., Mulder, van Maanen, & Forstmann, 2014). While these interpretations were based on theoretical assumptions rather than driven by data, in the current study, we assessed the exact cognitive stage that accounted for the differences. Interestingly, our results suggest that while stage 4 determines the familiarity of the words, this is only used as input for a decision process in stage 5.

The extension of ERPs, BRPs, showed that this decision process was located between two commonly identified ERPs: the N400 and P600. The larger N400 amplitudes for pseudowords compared to words in LD were previously explained through a lexical identification process (Meade, Grainger, & Holcomb, 2019). Because a lexical representation of pseudowords does not exist, the lexical identification process requires additional effort. These ideas can be integrated with the decisional account by assuming that the lexical identification process feeds into the decision. Here, the N400 might signify the starting point of the decision process – with different levels for the different conditions.

Additionally, the BRPs showed that not only the amplitude of the P600 differed between conditions but also its latency. While the P600 has been commonly found in language studies, there is an ongoing discussion on its interpretation and relation to another extensively studied positive-going component, the P300b (Polich, 2007). According to one of the views, these components are overlapping and share their functional interpretation (for an overview, see Leckey & Federmeier, 2020). Because the P300b is commonly associated with decision making and appears in binary decision tasks (Kutas, Van Petten, & Kluender, 2006; Twomey, Murphy, Kelly, & O’Connell, 2015), this fits with our explanation of Stage 5 as a decision stage. Note that there is an alternative explanation where the two components are assumed to be independent of each other with different neuronal generators (Frisch, Kotz, von Cramon, & Friederici, 2003).

Although in LD studies the N400 and P600 components were mostly reported in isolation, a recent study suggested an integrated functional interpretation of these components (Delogu, Brouwer, & Crocker, 2019). To this end, they were interpreted as part of a single process (labelled Retrieval-Integration account, see Brouwer, Fitz, & Hoeks, 2012) with the N400 indicating memory retrieval and the P600 semantic

integration. While this study established the importance of integration of these components, it is difficult to translate these findings on our data where the words were studied in isolation rather than in the context of sentences. That said, our study showed that these components should be studied together, as the latency between those components seems to be responsible for the overall RT effects in lexical decision.

Finally, there have been various approaches to model the word recognition process in LDT. In the past, DDM represented the standard approach to modeling data from this task (e.g., Donkin & Heathcote, 2009; Ratcliff et al., 2004; Tillman et al., 2017). While DDM allows decomposing data into decision and non-decision time, it does not lead to exactly the same conclusions as HsMM-MVPA does: where the HsMM-MVPA method attributed all condition effects to a single decision stage, some DDM results also indicate non-decision time differences (Donkin & Heathcote, 2009; Gomez & Perea, 2014). An important reason for that is the different sources of information that enter the model, namely, a single reaction time value per trial in the case of DDM and many EEG samples per trial in the case of HsMM-MVPA. While different variability in RTs per condition might lead to differences in non-decision time estimates in DDM modeling, the HsMM-MVPA method locates the same bumps on each trial, theoretically resulting in better stage length estimates. However, this is naturally dependent on the quality of the data, and EEG data might incorporate more noise than behavioral data. In our recent work, we discuss in more detail some differences in non-decision time estimations resulting from these methods (Berbery et al., 2021).

5. Limitation and future research

It should be noted that while both ERPs and HsMM-MVPA results indicated that the main difference between conditions started around 400 ms post-stimulus, the ERPs also showed slight differences between random non-words and other conditions starting around 200 ms at P8 and F7. However, these differences were limited to the amplitude of the signal, and were not observed in the temporal location of the peaks. The HsMM-MVPA method attempts to find bumps, and assumes no amplitude differences in the flats between consecutive bumps (Anderson et al., 2016; van Maanen et al., 2021). Given that the amplitude differences are also still visible in the BRPs (Fig. 7), this assumption is not fully met here. This suggests that there is a (small) processing difference for random strings, but that this processing difference does not result in a difference in stage topology or duration (see *Model selection*), at least not one that can be detected in the current dataset.

Additionally, HsMM-MVPA analysis relies on the assumption that cognitive stages are shared across participants and trials (van Maanen et al., 2021). While the current implementation of the method allows for flexibility in the duration of the stages, it is not aimed at exploring individual differences due to the absence of subject-specific parameters. We believe that the method could be further extended to include inter-individual variability. Such an extension would, for example, allow observing how the individual differences in lexical familiarity influence lexical decision.

While the current findings implicate differences between conditions in later stages, these stages might not be fixed and could be modulated by an attentional factor. In a recent study, a word frequency manipulation resulted in much earlier ERP differences in a go/no-go version of LDT than in a standard LDT (Vergara-Martínez, Gomez, & Perea, 2020). This was interpreted as relocation of attentional resources towards stimulus processing compared to a standard LDT (Smid, Fiedler, & Heinze, 2000). For HsMM-MVPA models it might mean that the observed difference in conditions would be present in stages other than decision, or that the decision stage would occur earlier. Altogether this highlights an important direction for a future HsMM-MVPA application.

6. Conclusion

In conclusion, we investigated cognitive processes underlying LD

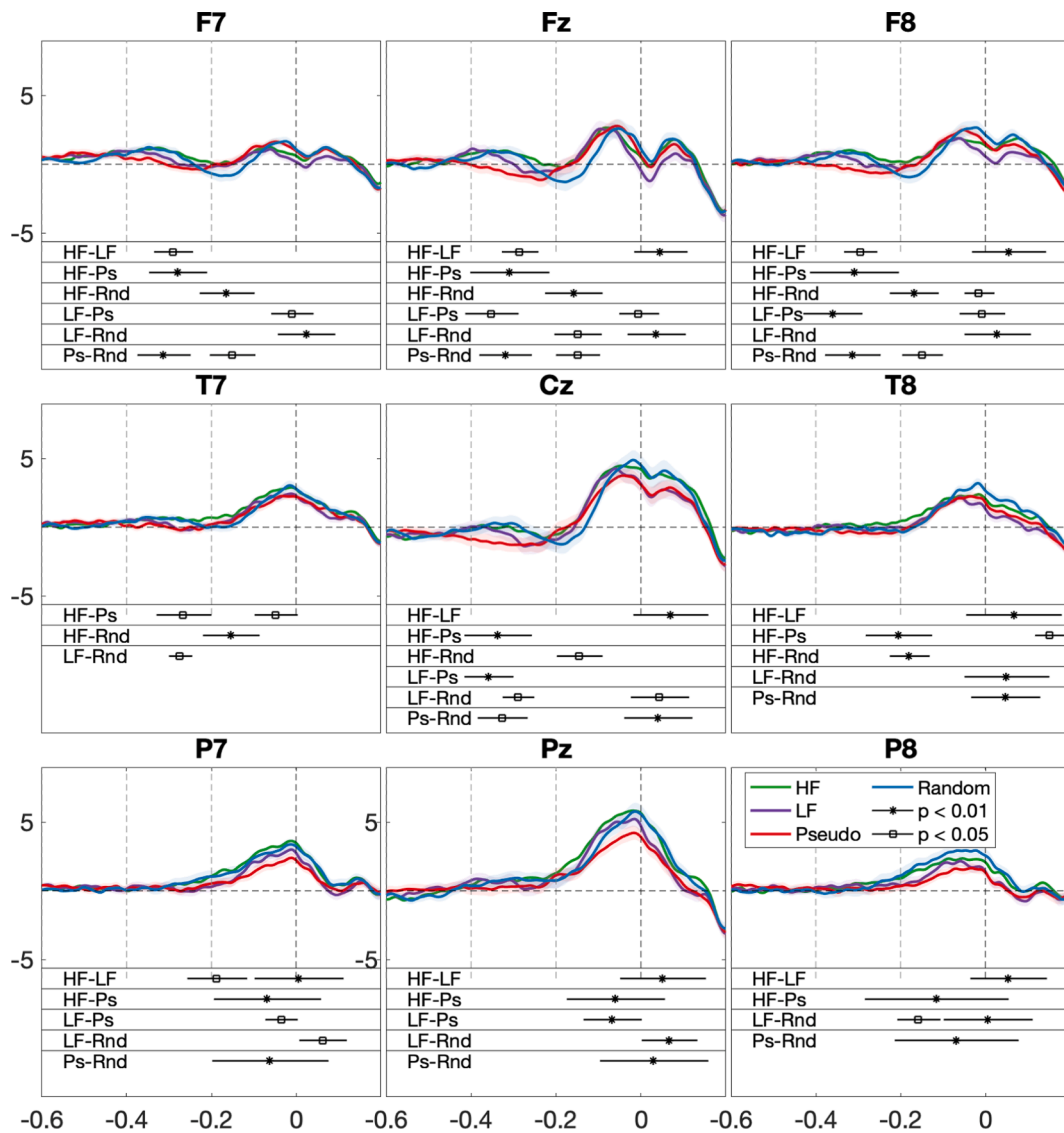


Fig. 8. Grand average response-locked ERPs for four conditions: HF (high frequency), LF (low frequency), Ps (pseudowords) and Rnd (random non-words) complemented with standard errors (shaded areas). The lines below grand averages indicate the significant clusters of differences between conditions. The location of these lines on horizontal axis corresponds to the temporal locations of the significant differences. The clusters significant at 0.05 are marked with asterisk (*) and at 0.01 with rectangle (□).

using a data-driven approach, namely, HsMM-MVPA applied to EEG data. We found that commonly reported main effects of word frequency and nonword lexicality were located in a single processing stage. The onset and termination of this processing stage overlapped with two language-related ERP components: the N400 and P600. Based on its temporal location and brain activation, this stage was interpreted as a decision process.

CRediT authorship contribution statement

Hermine S. Berberyan: Conceptualization, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Hedderik van Rijn:** Conceptualization, Funding acquisition, Methodology, Supervision, Writing – review & editing. **Jelmer P. Borst:** Conceptualization, Resources, Investigation, Methodology, Project administration, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial

interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to thank Rosa Verhoeven for her help with data collection and Oscar Portoles for his helpful suggestions on HsMM-MVPA analysis.

Funding sources

Van Rijn was supported by the research programme “Interval Timing in the Real World: A Functional, Computational and Neuroscience Approach,” with Project Number 453-16-005, financed by the Netherlands Organisation for Scientific Research (NWO).

Code availability

The code required to reproduce the analyses is available at <https://osf.io/z49me/>.

Appendix 1. Response-locked ERPs

See Fig. 8.

References

- Adams, M. J. (1979). Models of word recognition. *Cognitive Psychology*, 11(2), 133–176. [https://doi.org/10.1016/0010-0285\(79\)90008-2](https://doi.org/10.1016/0010-0285(79)90008-2).
- Aly, M., Yonelinas, A. P., Kishiyama, M. M., & Knight, R. T. (2011). Damage to the lateral prefrontal cortex impairs familiarity but not recollection. *Behavioural Brain Research*, 225(1), 297–304. <https://doi.org/10.1016/j.bbr.2011.07.043>.
- Anderson, J. R., Borst, J. P., Fincham, J. M., Ghuman, A. S., Tenison, C., & Zhang, Q. (2018). The common time course of memory processes revealed. *Psychological Science*, 29(9), 1463–1474. <https://doi.org/10.1177/0956797618774526>.
- Anderson, J. R., Zhang, Q., Borst, J. P., & Walsh, M. M. (2016). The discovery of processing stages: Extension of Sternberg's method. *Psychological Review*, 123(5), 481–509. <https://doi.org/10.1037/rev0000030>.
- Andrews, S. (1989). Frequency and neighborhood effects on lexical access: Activation or search? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(5), 802–814. <https://doi.org/10.1037/0278-7393.15.5.802>.
- Andrews, S., & Heathcote, A. (2001). Distinguishing common and task-specific processes in word identification: A matter of some moment? *Journal of Experimental Psychology: Learning Memory and Cognition*, 27(2), 514–544. <https://doi.org/10.1037/0278-7393.27.2.514>.
- Baayen, R. Harald, Piepenbrock, Richard, & van Rijn, Hedderik (1993). *The CELEX lexical data base on CD-ROM*. Philadelphia, PA: Linguistic Data Consortium.
- Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance*, 10(3), 340–357. <https://doi.org/10.1037/0096-1523.10.3.340>.
- Balota, D. A., & Spieler, D. H. (1999). Word frequency, repetition, and lexicality effects in word recognition tasks: Beyond measures of central tendency. *Journal of Experimental Psychology: General*, 128(1), 32–55. <https://doi.org/10.1037/0096-3445.128.1.32>.
- Barber, H. A., & Kutas, M. (2007). Interplay between computational models and cognitive electrophysiology in visual word recognition. *Brain Research Reviews*, 53(1), 98–123. <https://doi.org/10.1016/j.brainresrev.2006.07.002>.
- Barber, H. A., Otten, L. J., Kousta, S. T., & Vigliocco, G. (2013). Concreteness in word processing: ERP and behavioral effects in a lexical decision task. *Brain and Language*, 125(1), 47–53. <https://doi.org/10.1016/j.bandl.2013.01.005>.
- Barber, H., Vergara, M., & Carreiras, M. (2004). Syllable-frequency effects in visual word recognition: Evidence from ERPs. *NeuroReport*, 15(3), 545–548. <https://doi.org/10.1097/00001756-200403010-00032>.
- Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. ArXiv E-Print, under Revision. Retrieved from <http://arxiv.org/abs/1506.04967>.
- Bates, D. M., & DebRoy, S. (2004). Linear mixed models and penalized least squares. *Journal of Multivariate Analysis*, 91(1), 1–17. <https://doi.org/10.1016/j.jmva.2004.04.013>.
- Becker, C. A. (1980). Semantic context effects in visual word recognition: An analysis of semantic strategies. *Memory & Cognition*, 8(6), 493–512. <https://doi.org/10.3758/BF03213769>.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>.
- Berberyan, H. S., van Maanen, L., van Rijn, H., & Borst, J. (2021). EEG-based identification of evidence accumulation stages in decision-making. *Journal of Cognitive Neuroscience*, 33(3), 510–527. https://doi.org/10.1162/jocn_a.01663.
- Bermúdez-Margaretto, B., Beltrán, D., Domínguez, A., & Cuetos, F. (2015). Repeated exposure to “meaningless” pseudowords modulates LPC, but not N(FN)400. *Brain Topography*, 28(6), 838–851. <https://doi.org/10.1007/s10548-014-0403-5>.
- Besner, D. (1983). Basic decoding components in reading: Two dissociable feature extraction processes. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, 37(3), 429–438. <https://doi.org/10.1037/h0080739>.
- Borst, J. P., & Anderson, J. R. (in press). Discovering cognitive stages in M/EEG data to inform cognitive models. In B. U. Forstmann & B. M. Turner, (Eds.), *An introduction to model-based cognitive neuroscience*. Springer: New York.
- Borst, J. P., & Anderson, J. R. (2015). The discovery of processing stages: Analyzing EEG data with hidden semi-Markov models. *NeuroImage*, 108(1), 60–73. <https://doi.org/10.1016/j.neuroimage.2014.12.029>.
- Borst, J. P., Ghuman, A. S., & Anderson, J. R. (2016). Tracking cognitive processing stages with MEG: A spatio-temporal model of associative recognition in the brain. *NeuroImage*, 141, 416–430. <https://doi.org/10.1016/j.neuroimage.2016.08.002>.
- Braun, M., Jacobs, A. M., Hahne, A., Ricker, B., Hofmann, M., & Hutzler, F. (2006). Model-generated lexical activity predicts graded ERP amplitudes in lexical decision. *Brain Research*, 1073–1074, 431–439. <https://doi.org/10.1016/j.brainres.2005.12.078>.
- Brouwer, H., Fitz, H., & Hoeks, J. (2012). Getting real about Semantic Illusions: Rethinking the functional role of the P600 in language comprehension. *Brain Research*, 1446, 127–143. <https://doi.org/10.1016/j.brainres.2012.01.055>.
- Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect. *Experimental Psychology*, 58(5), 412–424. <https://doi.org/10.1027/1618-3169/a000123>.
- Carreiras, M., Mechelli, A., Estévez, A., & Price, C. J. (2007). Brain activation for lexical decision and reading aloud: Two sides of the same coin? *Journal of Cognitive Neuroscience*, 19(3), 433–444. <https://doi.org/10.1162/jocn.2007.19.3.433>.
- Carreiras, M., Vergara, M., & Barber, H. (2005). Early event-related potential effects of syllabic processing during visual word recognition. *Journal of Cognitive Neuroscience*, 17(11), 1803–1817. <https://doi.org/10.1162/089892905774589217>.
- Carreiras, M., Vergara, M., & Perea, M. (2007). ERP correlates of transposed-letter similarity effects: Are consonants processed differently from vowels? *Neuroscience Letters*, 419(3), 219–224. <https://doi.org/10.1016/j.neulet.2007.04.053>.
- Coch, D., & Mitra, P. (2010). Word and pseudoword superiority effects reflected in the ERP waveform. *Brain Research*, 1329, 159–174. <https://doi.org/10.1016/j.brainres.2010.02.084>.
- Cohen, M. X. (2019). Analyzing neural time series data. *Analyzing Neural Time Series Data*. <https://doi.org/10.7551/mitpress/9609.001.0001>.
- Curran, T. (1999). The electrophysiology of incidental and intentional retrieval: ERP old/new effects in lexical decision and recognition memory. *Neuropsychologia*, 37(7), 771–785. [https://doi.org/10.1016/S0028-3932\(98\)00133-X](https://doi.org/10.1016/S0028-3932(98)00133-X).
- Delogu, F., Brouwer, H., & Crocker, M. W. (2019). Event-related potentials index lexical retrieval (N400) and integration (P600) during language comprehension. *Brain and Cognition*, 135, Article 103569. <https://doi.org/10.1016/j.bandc.2019.05.007>.
- Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, 134(1), 9–21. <https://doi.org/10.1016/j.jneumeth.2003.10.009>.
- Donkin, C., & Heathcote, A. (2009). Non-decision time effects in the lexical decision task. *Proceedings of the 31st annual conference of the cognitive science society*. Austin: Cognitive Science Society.
- Eichenbaum, H., Yonelinas, A. P., & Ranganath, C. (2007). The medial temporal lobe and recognition memory. *Annual Review of Neuroscience*, 30(1), 123–152. <https://doi.org/10.1146/annurev.neuro.30.051606.094328>.
- Fiez, J. A., Balota, D. A., Raichle, M. E., & Petersen, S. E. (1999). Effects of lexicality, frequency, and spelling-to-sound consistency on the functional anatomy of reading. *Neuron*, 24(1), 205–218. [https://doi.org/10.1016/S0896-6273\(00\)80833-8](https://doi.org/10.1016/S0896-6273(00)80833-8).
- Frisch, S., Kotz, S. A., von Cramon, D. Y., & Friederici, A. D. (2003). Why the P600 is not just a P300: The role of the basal ganglia. *Clinical Neurophysiology*, 114(2), 336–340. [https://doi.org/10.1016/S1388-2457\(02\)00366-8](https://doi.org/10.1016/S1388-2457(02)00366-8).
- Gomez, P., & Perea, M. (2014). Decomposing encoding and decisional components in visual-word recognition: A diffusion model analysis. *Quarterly Journal of Experimental Psychology*, 67(12), 2455–2466. <https://doi.org/10.1080/17470218.2014.937447>.
- Grainger, J. (1990). Word frequency and neighborhood frequency effects in lexical decision and naming. *Journal of Memory and Language*, 29(2), 228–244. [https://doi.org/10.1016/0749-596X\(90\)90074-A](https://doi.org/10.1016/0749-596X(90)90074-A).
- Grainger, J., Bouttevin, S., Truc, C., Bastien, M., & Ziegler, J. (2003). Word superiority, pseudoword superiority, and learning to read: A comparison of dyslexic and normal readers. *Brain and Language*, 87(3), 432–440. [https://doi.org/10.1016/S0093-934X\(03\)00145-7](https://doi.org/10.1016/S0093-934X(03)00145-7).
- Hintzman, D. L., & Curran, T. (1997). Comparing retrieval dynamics in recognition memory and lexical decision. *Journal of Experimental Psychology: General*, 126(3), 228–247. <https://doi.org/10.1037/0096-3445.126.3.228>.
- Holcomb, P. J., Grainger, J., & O'Rourke, T. (2002). An electrophysiological study of the effects of orthographic neighborhood size on printed word perception. *Journal of Cognitive Neuroscience*, 14(6), 938–950. <https://doi.org/10.1162/089892902760191153>.
- Holcomb, P. J., & Neville, H. J. (1990). Auditory and visual semantic priming in lexical decision: A comparison using event-related brain potentials. *Language and Cognitive Processes*, 5(4), 281–312. <https://doi.org/10.1080/01690969008407065>.
- Horn, M., Jardri, R., D'Hondt, F., Vaiva, G., Thomas, P., & Pins, D. (2016). The multiple neural networks of familiarity: A meta-analysis of functional imaging studies. *Cognitive, Affective, & Behavioral Neuroscience*, 16(1), 176–190. <https://doi.org/10.3758/s13415-015-0392-1>.
- Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono- and disyllabic words and nonwords. *Frontiers in Psychology*, 1(NOV). <https://doi.org/10.3389/fpsyg.2010.00174>.
- Kok, A. (2001). On the utility of P3 amplitude as a measure of processing capacity. *Psychophysiology*, 38(3). <https://doi.org/10.1017/S0048577201990559>.
- Kutas, M., & 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*, 62(1), 621–647. <https://doi.org/10.1146/annurev.psych.093008.131123>.
- Kutas, M., Van Petten, C. K., & Kluender, R. (2006). Psycholinguistics electrified II (1994–2005). In *Handbook of psycholinguistics* (pp. 659–724). Elsevier. 10.1016/B978-012369374-7/50018-3.
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13). <https://doi.org/10.18637/jss.v082.i13>.
- Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (de)constructing the N400. *Nature Reviews Neuroscience*, 9(12), 920–933. <https://doi.org/10.1038/nrn2532>.
- Leckey, M., & Federmeier, K. D. (2020). The P3b and P600(s): Positive contributions to language comprehension. *Psychophysiology*, 57(7). <https://doi.org/10.1111/psyp.13351>.
- Makeig, S., Westerfield, M., Jung, T. P., Enghoff, S., Townsend, J., Courchesne, E., & Sejnowski, T. J. (2002). Dynamic brain sources of visual evoked responses. *Science*, 295(5555), 690–694. <https://doi.org/10.1126/science.1066168>.

- Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, 164(1), 177–190. <https://doi.org/10.1016/j.jneumeth.2007.03.024>.
- Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*. <https://doi.org/10.3758/s13428-011-0168-7>.
- McCann, R. S., Besner, D., & Davelaar, E. (1988). Word recognition and identification: Do word-frequency effects reflect lexical access? *Journal of Experimental Psychology: Human Perception and Performance*, 14(4), 693–706. <https://doi.org/10.1037/0096-1523.14.4.693>.
- Meade, G., Grainger, J., & Holcomb, P. J. (2019). Task modulates ERP effects of orthographic neighborhood for pseudowords but not words. *Neuropsychologia*, 129, 385–396. <https://doi.org/10.1016/j.neuropsychologia.2019.02.014>.
- Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 90(2), 227–234. <https://doi.org/10.1037/h0031564>.
- Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology*, 4(2), 61–64. <https://doi.org/10.20982/tqmp.04.2.p061>.
- Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, 76(2), 165–178. <https://doi.org/10.1037/h0027366>.
- Mulder, M. J., van Maanen, L., & Forstmann, B. U. (2014). Perceptual decision neurosciences - a model-based review. *Neuroscience*. <https://doi.org/10.1016/j.neuroscience.2014.07.031>.
- Murphy, K. A., Jogia, J., & Talcott, J. B. (2019). On the neural basis of word reading: A meta-analysis of fMRI evidence using activation likelihood estimation. *Journal of Neurolinguistics*, 49, 71–83. <https://doi.org/10.1016/j.jneuroling.2018.08.005>.
- Nieuwenhuis, S., Aston-Jones, G., & Cohen, J. D. (2005). Decision making, the P3, and the locus coeruleus–norepinephrine system. *Psychological Bulletin*, 131(4), 510–532. <https://doi.org/10.1037/0033-2909.131.4.510>.
- Nobre, A. C., & McCarthy, G. (1994). Language-related ERPs: Scalp distributions and modulation by word type and semantic priming. *Journal of Cognitive Neuroscience*, 6(3), 233–255. <https://doi.org/10.1162/jocn.1994.6.3.233>.
- Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological Review*, 110(4), 611–646. <https://doi.org/10.1037/0033-295X.110.4.611>.
- Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, 2011, 1–9. <https://doi.org/10.1155/2011/156869>.
- Paap, K. R., Newsome, S. L., McDonald, J. E., & Schvaneveldt, R. W. (1982). An activation-verification model for letter and word recognition: The word-superiority effect. *Psychological Review*, 89(5), 573–594. <https://doi.org/10.1037/0033-295X.89.5.573>.
- Perea, M., Rosa, E., & Gómez, C. (2002). Is the go/no-go lexical decision task an alternative to the yes/no lexical decision task? *Memory & Cognition*, 30(1), 34–45. <https://doi.org/10.3758/BF03195263>.
- Perea, M., Rosa, E., & Gómez, C. (2005). The frequency effect for pseudowords in the lexical decision task. *Perception & Psychophysics*, 67(2), 301–314. <https://doi.org/10.3758/BF03206493>.
- Plaut, D. C. (1997). Structure and function in the lexical system: Insights from distributed models of word reading and lexical decision. *Language and Cognitive Processes*, 12(5–6), 765–806. <https://doi.org/10.1080/016909697386682>.
- Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology*, 118(10), 2128–2148. <https://doi.org/10.1016/j.clinph.2007.04.019>.
- Polich, J., & Donchin, E. (1988). P300 and the word frequency effect. *Electroencephalography and Clinical Neurophysiology*, 70(1), 33–45. [https://doi.org/10.1016/0013-4694\(88\)90192-7](https://doi.org/10.1016/0013-4694(88)90192-7).
- Portoles, O., Borst, J. P., & van Vugt, M. K. (2018). Characterizing synchrony patterns across cognitive task stages of associative recognition memory. *European Journal of Neuroscience*, 48(8), 2759–2769. <https://doi.org/10.1111/ejn.13817>.
- Proverbio, A. M., & Adorni, R. (2008). Orthographic familiarity, phonological legality and number of orthographic neighbours affect the onset of ERP lexical effects. *Behavioral and Brain Functions*, 4(1), 27. <https://doi.org/10.1186/1744-9081-4-27>.
- Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 257–286. <https://doi.org/10.1109/5.18626>.
- Ratcliff, R., McKoon, G., & Gomez, P. (2004). A diffusion model account of the lexical decision task. *Psychological Review*, 111(1), 159–182. <https://doi.org/10.1037/0033-295X.111.1.159>.
- Rugg, M. D. (1990). Event-related brain potentials dissociate repetition effects of high- and low-frequency words. *Memory & Cognition*, 18(4), 367–379. <https://doi.org/10.3758/BF03197126>.
- Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance*, 3(1), 1–17. <https://doi.org/10.1037/0096-1523.3.1.1>.
- Schilling, H. E. H., Rayner, K., & Chumbley, J. I. (1998). Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. *Memory and Cognition*, 26(6), 1270–1281. <https://doi.org/10.3758/BF03201199>.
- Schroeder, C. E., Steinschneider, M., Javitt, D. C., Tenke, C. E., Givre, S. J., Mehta, A. D., ... Vaughan, H. G. (1995). Localization of ERP generators and identification of underlying neural processes. *Electroencephalography and Clinical Neurophysiology Supplement*, 44, 55–75. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/7649056>.
- Seidenberg, Mark S., & McClelland, James L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, 96(4), 523–568. <https://doi.org/10.1037/0033-295X.96.4.523>.
- Shah, A. S., Bressler, S. L., Knuth, K. H., Ding, M., Mehta, A. D., Ulbert, I., & Schroeder, C. E. (2004). Neural dynamics and the fundamental mechanisms of event-related brain potentials. *Cerebral Cortex*, 14(5), 476–483. <https://doi.org/10.1093/cercor/bhh009>.
- Smid, H. G. O. M., Fiedler, R., & Heinze, H.-J. (2000). An electrophysiological study of the insertion of overt response choice. *Journal of Experimental Psychology: Human Perception and Performance*, 26(3), 1053–1071. <https://doi.org/10.1037/0096-1523.26.3.1053>.
- Soltani, M., & Knight, R. T. (2000). Neural Origins of the P300. *Critical Reviews™ in Neurobiology*, 14(3–4), 26. <https://doi.org/10.1615/CritRevNeurobiol.v14.i3-4.20>.
- Stone, G. O., & Van Orden, G. C. (1993). Strategic control of processing in word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 19(4), 744–774. <https://doi.org/10.1037/0096-1523.19.4.744>.
- Sutton, S., Ruchkin, D. S., Munson, R., Kietzman, M. L., & Hammer, M. (1982). Event-related potentials in a two-interval forced-choice detection task. *Perception & Psychophysics*, 32(4), 360–374. <https://doi.org/10.3758/BF03206242>.
- Taylor, S. F., Stern, E. R., & Gehring, W. J. (2007). Neural systems for error monitoring. *The Neuroscientist*, 13(2), 160–172. <https://doi.org/10.1177/1073858406298184>.
- Tillman, G., Osth, A. F., van Ravenzwaaj, D., & Heathcote, A. (2017). A diffusion decision model analysis of evidence variability in the lexical decision task. *Psychonomic Bulletin & Review*, 24(6), 1949–1956. <https://doi.org/10.3758/s13423-017-1259-y>.
- Twomey, D. M., Murphy, P. R., Kelly, S. P., & O'Connell, R. G. (2015). The classic P300 encodes a build-to-threshold decision variable. *European Journal of Neuroscience*, 42(1), 1636–1643. <https://doi.org/10.1111/ejn.12936>.
- van Maanen, L., Portoles, O., & Borst, J. P. (2021). The discovery and interpretation of evidence accumulation stages. *Computational Brain & Behavior*. <https://doi.org/10.1007/s42113-021-00105-2>.
- Vergara-Martínez, M., Gomez, P., & Perea, M. (2020). Should I stay or should I go? An ERP analysis of two-choice versus go/no-go response procedures in lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(11), 2034–2048. <https://doi.org/10.1037/xlm0000942>.
- Wagenmakers, E. J., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language*, 58(1), 140–159. <https://doi.org/10.1016/j.jml.2007.04.006>.
- Wagenmakers, E. J., Steyvers, M., Raaijmakers, J. G. W., Shiffrin, R. M., van Rijn, H., & Zeelenberg, R. (2004). A model for evidence accumulation in the lexical decision task. *Cognitive Psychology*, 48(3), 332–367. <https://doi.org/10.1016/j.cogpsych.2003.08.001>.
- Walsh, M. M., Gunzelmann, G., & Anderson, J. R. (2017). Relationship of P3b single-trial latencies and response times in one, two, and three-stimulus oddball tasks. *Biological Psychology*, 123, 47–61. <https://doi.org/10.1016/j.biopsycho.2016.11.011>.
- Wang, Q., & Yuan, J. (2008). N400 lexicality effect in highly blurred Chinese words: Evidence for automatic processing. *NeuroReport*, 19(2), 173–178. <https://doi.org/10.1097/WNR.0b013e3282f3e3f2>.
- Wang, X., Yang, J., Yang, J., Mencl, W. E., Shu, H., & Zevin, J. D. (2015). Language differences in the brain network for reading in naturalistic story reading and lexical decision. *PLOS ONE*, 10(5), Article e0124388. <https://doi.org/10.1371/journal.pone.0124388>.
- Woodruff, C. C., Hayama, H. R., & Rugg, M. D. (2006). Electrophysiological dissociation of the neural correlates of recollection and familiarity. *Brain Research*, 1100(1), 125–135. <https://doi.org/10.1016/j.brainres.2006.05.019>.
- Young, M. P., & Rugg, M. D. (1992). Word frequency and multiple repetition as determinants of the modulation of event-related potentials in a semantic classification task. *Psychophysiology*, 29(6), 664–676. <https://doi.org/10.1111/j.1469-8986.1992.tb02044.x>.
- Yu, S.-Z. (2010). Hidden semi-Markov models. *Artificial Intelligence*, 174(2), 215–243. <https://doi.org/10.1016/j.artint.2009.11.011>.
- Zeelenberg, R., Wagenmakers, E.-J., & Shiffrin, R. M. (2004). Nonword repetition priming in lexical decision reverses as a function of study task and speed stress. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(1), 270–277. <https://doi.org/10.1037/0278-7393.30.1.270>.
- Zhang, Q., Borst, J. P., Kass, R. E., & Anderson, J. R. (2017). Inter-subject alignment of MEG datasets in a common representational space. *Human Brain Mapping*, 38(9), 4287–4301. <https://doi.org/10.1002/hbm.23689>.
- Zhang, Q., van Vugt, M., Borst, J. P., & Anderson, J. R. (2018). Mapping working memory retrieval in space and in time: A combined electroencephalography and electrocorticography approach. *NeuroImage*, 174, 472–484. <https://doi.org/10.1016/j.neuroimage.2018.03.039>.
- Zhang, Q., Walsh, M. M., & Anderson, J. R. (2017). The effects of probe similarity on retrieval and comparison processes in associative recognition. *Journal of Cognitive Neuroscience*, 29(2), 352–367. <https://doi.org/10.1162/jocn.a.01059>.
- Zhang, Q., Walsh, M. M., & Anderson, J. R. (2018). The impact of inserting an additional mental process. *Computational Brain & Behavior*, 1(1), 22–35. <https://doi.org/10.1007/s42113-018-0002-8>.
- Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, 20(1), 45–57. <https://doi.org/10.1109/42.906424>.
- Ziegler, J. C., Besson, M., Jacobs, A. M., Nazir, T. A., & Carr, T. H. (1997). Word, pseudoword, and nonword processing: A multitask comparison using event-related brain potentials. *Journal of Cognitive Neuroscience*, 9(6), 758–775. <https://doi.org/10.1162/jocn.1997.9.6.758>.