SCIENTIFIC REPORTS
nature research

OPEN

Electrophysiological responses of relatedness to consecutive word stimuli in relation to an actively recollected target word

Received: 21 December 2018
Accepted: 19 September 2019
Published online: 10 October 2019

Karen Dijkstra, Jason Farquhar & Peter Desain

In this paper, we investigate the robustness of electrophysiological responses of relatedness to multiple consecutive word stimuli (probes), in relation to an actively recollected target word. Such relatedness information could be used by a Brain Computer Interface to infer the active semantic concept on a user's mind, by integrating the knowledge of the relationship between the multiple probe words and the 'unknown' target. Such a BCI can take advantage of the N400: an event related potential that is sensitive to semantic content of a stimulus in relation to an established semantic context. However, it is unknown whether the N400 is suited for the multiple probing paradigm we propose, as other intervening words might distract from the established context (i.e., the target word). We perform an experiment in which we present up to ten words after an initial target word, and find no attenuation of the strength of the N400 in grand average ERPs and no decrease in classification accuracy for probes occurring later in the sequences. These results are groundwork for developing a BCI that infers the concept on a user's mind through repeated probing, however, low single trial decoding accuracy, and high subject variability may limit practical applicability.

Brain Computer Interfaces (BCIs) use brain activity as a direct input for a computer. Many BCI applications are designed to offer alternative means of communication to people that are no longer able to use conventional input devices (e.g., keyboards). Existing communication BCIs come in a number of paradigms, achieving communication in different ways. Some offer binary choices selections, that, for instance, allow for "Yes"/"No" answers to questions (e.g.,[1–3]). Others allow users to spell messages, by selecting characters one by one[4–7]. Each approach has their strengths and weaknesses, allowing the choice of BCI to vary depending on the needs and preferences of a user.

Here, we are working towards a BCI that allows for the selection of words as a unit. In an approach similar to the game '20 questions', in which someone tries to guess the person or object that someone is thinking of by asking only yes-no questions, we envision a BCI that infers the concept on a user's mind by, in essence, asking them if a presented word is related to the to-be-inferred concept. More specifically, the proposed BCI presents a word, uses the electrophysiological responses of semantic relatedness to infer this word's relation to the target concept, and updates its belief state based on this evidence. We refer to these words, designed to elicit information about the unknown target, as 'probe' words. The BCI continues by presenting a new probe word, measuring another brain response, and so on, until it has sufficient confidence that the target concept has been identified. This can be employed as a word selection application, for purposes of communication (e.g., in sentence generation, or perhaps topic selection) as an alternative for existing BCI communication approaches, or for tip-of-the-tongue scenarios, in which the word for a to-be-communicated concept cannot be retrieved, which can be a problem for patients with anomic aphasia[8], for instance. In the latter case, these patients could benefit from an application that can help them find their word when issues arise, though currently, even state-of-the-art BCIs are hard to incorporate into daily life, making this an unrealistic application in the short term.

Such an approach toward word selection has been outlined in the past[9]. In that study by Geuze *et al.* (2014), users encoded a probe's relatedness status using deliberate responses: Users were presented with a stream of probes and asked to press a button when a presented word was related; a task designed to simultaneously induce a movement related de-synchronization (ERD) and elicit a P300 for related probes (due to their explicit task-related nature).

Radboud University, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, Netherlands. Correspondence and requests for materials should be addressed to K.D. (email: k.dijkstra@donders.ru.nl)

While using such deliberate signals is one way to approach such a BCI, there already exists a brain response that is inherently sensitive to the semantic content of a stimulus: the N400. It is an Event Related Potential (ERP) characterised as a wave that is more negative when a presented stimulus is unrelated, compared to a related stimulus, in relation to a previously established context[10]. Over the years it has proven a robust effect that can be elicited in a sentence context, in which the response to a word is measured based on its relation to earlier parts of the sentence, or in a word-pair context, in which a prime is presented followed by a related or unrelated probe. It is not only elicited with text on a screen, but also occurs when stimuli are pictures[11] or when presented auditorily, as speech[12,13], though the N400's scalp distribution can vary. There are also non-semantic factors that affect the N400 amplitude, for instance, the lexical properties of the presented stimulus word (e.g., the word's frequency)[14,15], or the repetition of a stimulus within in a short timeframe[16]. These are of less inherent interest for a BCI application, but may need to be accounted for as potential confounds.

Importantly, it is not necessary that subjects are explicitly tasked to evaluate the semantic content of a stimulus to elicit an N400, though it appears to require that a stimulus' meaning is processed (i.e., in a task where only the length of the word is relevant, no semantic priming effects from previous stimuli occur;[17,18]). Using such an "automatic" brain response might be more pleasant for the user, as they have to provide less effort in deliberately evoking the right response. However, if we want to use the N400 for this BCI paradigm, we first have to establish whether this signal lends itself to such an approach.

Van Vliet *et al.*[19], recognised the potential of the N400 for BCI purposes, and demonstrated that a prime word need not be presented, but can simply be recalled, to elicit an N400 to subsequent'probe' stimuli. To use for BCI purposes, the N400 should be detectable from a brain response to a single stimulus presentation, and not merely be observable in averages across or within participants. More recent work has shown that indeed the N400 can be detected from single word pair presentations (i.e., a target word and a single probe), with classification rates across individuals ranging from 54% to 67%[20].

Secondly, to infer the concept on a user's mind, a system would have to present a number of probe words consecutively, for the same concept, in order to collect sufficient evidence. Such a sequence of probe words could conceivably influence the context on a users mind, and it is therefore possible that results from a prime-single probe paradigm do not generalise to a setup with multiple (i.e., many) probes. In other words, we want to know whether single trial classification of relatedness based on the N400 is still possible when multiple consecutive probes are presented following an initial target word.

It is good to note that, given the described automatic nature of the N400, we might expect that in the Geuze *et al.* (2014) study[9], that focused on deliberate responses, to also elicit an N400 response. However, this N400 would predict a difference in the same direction as their expected P300: a more positive response for related probes after 300 ms, so it is hard to estimate the potential contribution of the N400 to their ERP findings. Notably, their grand average ERPs do not show the downward deflection (with a minimum near 400 ms) that is characteristic for the N400.

To establish what the effects of multiple consecutive probes are on the decoding of semantic relatedness, we designed an experiment in which we present a target word for participants to actively keep in mind, followed by up to ten probe words that are either related or unrelated to this target. Subjects are tasked to mentally evaluate the relatedness of probes, but without putting emphasis on related over unrelated probes. While presenting ten probes is not necessarily sufficient for decoding purposes, we will use this data to determine if there is any indication of attenuation (i.e., a reduction of the magnitude) of the N400, when comparing first probes and probes late in the sequence. Specifically, we will compare grand average ERPs in response to probes appearing immediately after the target, with those in response to probes at the end of the sequence, combining data across participants to increase the sensitivity for detecting a difference.

In addition to this, there is some evidence that tasks designed to elicit N400s can also induce changes in oscillatory brain responses[21]. To evaluate this we will analyse the data in the frequency domain to determine if there are any spectral differences that could be used as additional features for decoding semantic relatedness.

Results

Behavioural results. In each trial, participants were presented with a target word to keep in mind, followed by up to ten probe words. At the end of each trial, participants indicated with a button-press whether the most recently presented probe had been related or unrelated to the initial target word (left vs. right button-press respectively). We compared the answers of each participant to the label of that probe and calculated the percentage of agreeing answers across all trials. Agreement between the response and label ranged from 70% to 95% across participants (mean 87%). A mismatch between the response and label can reflect a user error, or a disagreement on the relatedness status of the probe, as relatedness judgements can vary between individuals. Average reaction times per participant ranged from 368 ms to 1420 ms (mean 632 ms), with a strong correlation between reaction time and age ($r = 0.80$, $n = 20$, $p = 0.00003$, ages 18–61, mean = 29, sd = 12). On average, reaction times to related probes were somewhat faster than to unrelated probes (578 ms and 650 ms respectively; agreeing responses only).

Initial ERP analysis. To determine if there were any differences in brain responses to related compared to unrelated probes, we computed a grand average ERP, averaging across all related and unrelated probes presented in the experiment. This analysis produced an unexpected result: the grand average showed a difference in response between related and unrelated probes *prior to* stimulus onset, with the ERP for unrelated probes being more negative before, at, and some period after time 0. A difference in ERPs occurring prior to presentation of the stimulus of interest cannot reflect a brain response to that stimulus, suggesting that there is some structural factor (e.g., a previous stimulus) affecting the responses, that did not average out in the ERP. We hypothesise that this is a consequence of our design: while two related stimuli were never presented sequentially, unrelated stimuli could appear after either a related or an unrelated probe. If we split the data into these three categories and compute

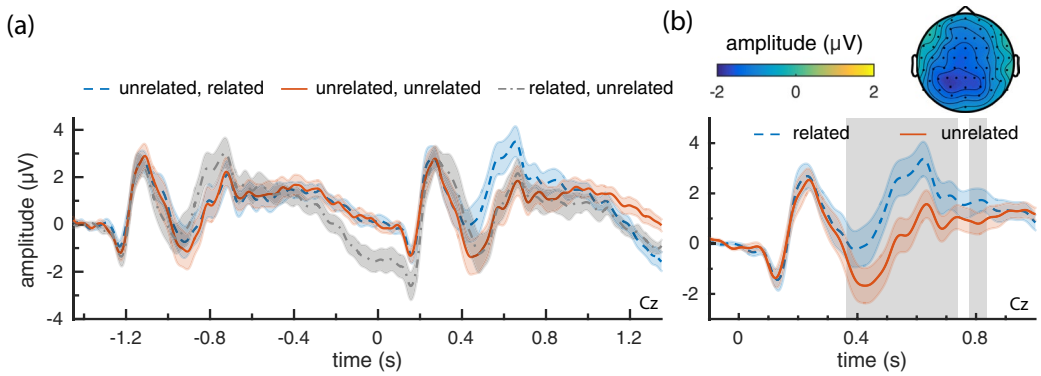

Figure 1. Grand average ERPs of responses to related and unrelated probes (Cz). (**a**) Responses to probes differentiated by their predecessor (related or unrelated; *'related, related'* did not occur in the experiment) from −1.35 s (presentation of previous probe) to 1.35 s (presentation of next probe). (**b**) Grand average of related and unrelated probes that followed an unrelated probe. The grey area marks, for this channel, the timepoints that were part of the significant cluster identified by the cluster permutation test. The accompanying topoplot from 300 to 800 ms has been included on top (unrelated-related).

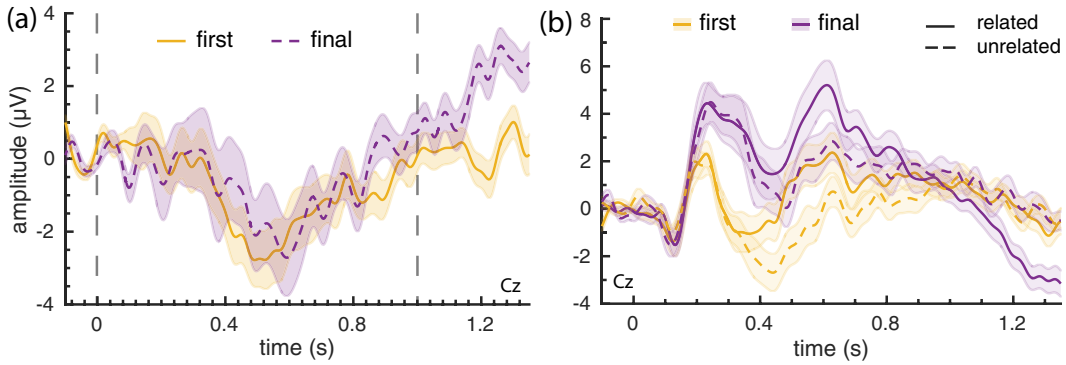

Figure 2. Grand average ERPs for probes presented immediately after a target (first; 1st position), and probes presented at the end of a trial sequence (final; 9th and 10th position) (Cz). (**a**) Difference waves of probe responses: unrelated minus related. No significant clusters were found in a cluster permutation test applied to the period from 0 to 1 s (indicated by the gray frame). (**b**) Related and unrelated ERPs from first and final probe positions.

grand averages we obtain the results in Fig. 1a (for electrode Cz; baselined to the predecessor stimulus). Here we see that the pre-stimulus difference occurs only for unrelated probes occurring after a related probe.

More specifically, there appears to be a late response that differs in amplitude between related and unrelated probes, which is still present when the next stimulus is presented. This is followed by a stimulus response (around 200 ms) after which the *related → unrelated* category behaves similarly to the *unrelated → unrelated*, until at least 1 s after stimulus presentation.

With responses to probes preceded by an unrelated probe appearing unconfounded (i.e., behaving more or less identically at stimulus onset), from here on we consider only these probe's data, excluding the *related → unrelated* data (unless noted otherwise), to ensure that effects are not an artifact of our experimental design.

ERP analyses. The resulting grand average ERP can be found in Fig. 1b. Here we observe a difference between the two conditions from around 400 ms to 800 ms, where the response to unrelated probes is more negative than to related probes, as would be expected for an N400. This difference corresponds to the significant cluster found in a cluster permutation test[22] ($p = 0.001$, two-tailed test, $\alpha = 0.05$, $n = 20$, cluster marked in grey), confirming that there is a significant difference in the response to related and unrelated probes.

In order to determine whether or not the magnitude of the ERP decreases after more probes have been presented, we compare the ERP in response to the first probe in each plot, to the ERPs of those in final positions (i.e., the 9th and 10th position). Note, that due to the exclusion of unrelated probes that were preceded by a related probe, there are fewer instances available for unrelated probes in the final positions (104 in the first and 53 in the final position respectively, per subject). For both locations (first vs final), we plot the difference ERP (unrelated-related) in Fig. 2a.

For both probes in the first and final positions, the response to unrelated probes was more negative than the response to related probes, resulting in a negative difference from around 400 ms to 800 ms. In the cluster

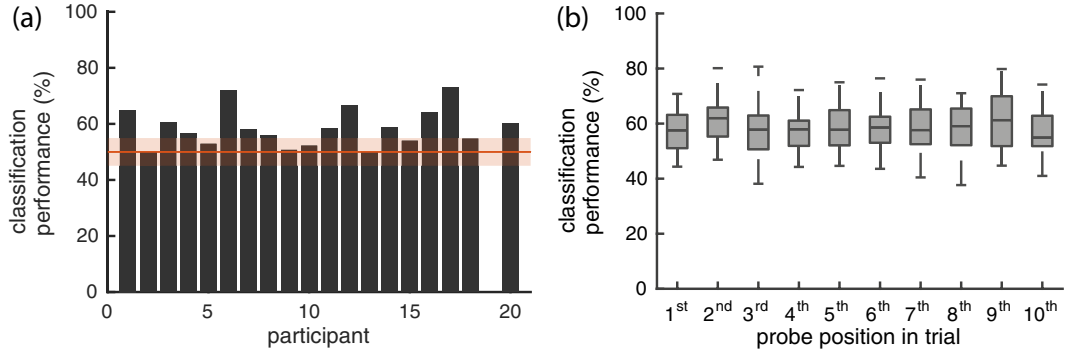

Figure 3. Classification accuracy of predicting (un)relatedness. (**a**) Classification accuracy per subject (balanced accuracy). Shaded in red, the 99.74% binomial confidence interval around 50% accuracy (interval: [0.4506, 0.5494]; Bonferroni-corrected for 19 subjects). (**b**) Classification accuracies for each position of a probe in the trial, across participants.

permutation applied to the two difference waves (from 0–1 s; indicated by the grey frame), no significant clusters were found (lowest cluster p-value observed: $p = 0.21$, two-tailed test, $\alpha = 0.05$, $n = 20$). Unexpectedly, the figure suggests that there is a difference after one second. To look at this in more detail, we plot the individual ERPs for related and unrelated for both the first and final positions in Fig. 2b. Here, we see that ERPs in response to first and final probes look different, even though their difference waves are similar. Furthermore, the positive difference wave after 1 s in the final position, appears to be due to a late negative response for related probes in particular. This may be related to the late negative wave we observed in Fig. 1a for related probes, which would imply this late negative wave did not occur for probes in the first position.

Predicting relatedness. As our interest is ultimately in the decoding of these signals from single probe presentations, we trained classifiers to predict for a given probe response whether it is related or unrelated, separately for each subject, using a leave-one-sequence-out crossvalidation approach. No classifier was trained for subject 19, as they were unable to complete the full experiment. The accuracy per subject can be found in Fig. 3a. Here, the accuracy of predicting whether a presented probe was related or unrelated is plotted per subject. To account for the larger number of unrelated probes in the test data, we report a balanced accuracy ((*sensitivity* + *specificity*)/2), rather than the percentage of correctly predicted probes. In the figure, we shade the 99.74% binomial confidence interval of chance performance in red to get an indication of which participant's results did not differ from chance (i.e., a 95% confidence interval, Bonferroni-corrected for 19 subjects, around 50%, estimated by drawing binomial proportions for each class to account for imbalance in class sizes). For 12 out of 19 participants, the classification accuracy could be distinguished from chance level. On average, the relatedness of probes was predicted correctly for 58.3% of probes (sd = 6.6%, AUC mean = 0.62, sd = 0.09). To ensure our crossvalidation approach did not overestimate the classification accuracies, we compare them to the accuracies obtained with a train-test split (80–20%). For this train-test split, the average accuracy across participants was 59.5%.

ERP plots for each individual subject are included in the Supplementary Information (Supplemental Fig. S1), with their respective classification accuracy, as these may provide an idea of the variability across subjects in the ERPs themselves.

Using the behavioural measure, we can determine whether there is a relation between how well the participant's behavioural responses agreed with our labels, and how well the relatedness of a probe can be predicted from their brain signals. The correlation between behavioural agreement and classifier accuracy, across participants, was moderately high ($r = 0.58$, $p = 0.009$, $\alpha = 0.05$).

In Fig. 3b we plot the classification accuracy based on where the probe occurred in a trial (i.e., the probe position). The boxplots show the distribution of performance across participants for a given probe position. There is no clear pattern visible between probe position and the accuracy of the prediction. To determine statistical (in) significance of any trend, we performed a permutation test, permuting the order of probe positions. Specifically, we compared the regression coefficient (of a line-fit) of the observed result against regression coefficients of results in which the order of probe positions was randomly permuted per subject (1000 permutations). The observed regression coefficient did not differ statistically from the coefficients from permuted results ($p = 0.562$; two-tailed alpha = 0.05).

Time frequency analysis. To determine if there was any difference in oscillatory brain activity when contrasting the related or unrelated probes, we also analysed the data in the time-frequency domain. A cluster permutation test applied to the Time Frequency Representations (TFRs) of related and unrelated probes, found a significant difference between the two conditions, identifying a significant negative cluster ($p = 0.001$, two-tailed, alpha = 0.05, $n = 20$)). To visualise this data, we obtain the difference between the the TFR for related probes and the TFR for unrelated probes (i.e., unrelated–related), as a fraction of the total power per frequency band (summed across the time-dimension) and plot this normalised difference, averaged across electrodes, in Fig. 4a. Clusters from a cluster-based permutation test are of limited use for determining which frequencies, timepoints, and electrodes in particular contribute to the significant difference, as the test does not control for the false alarm

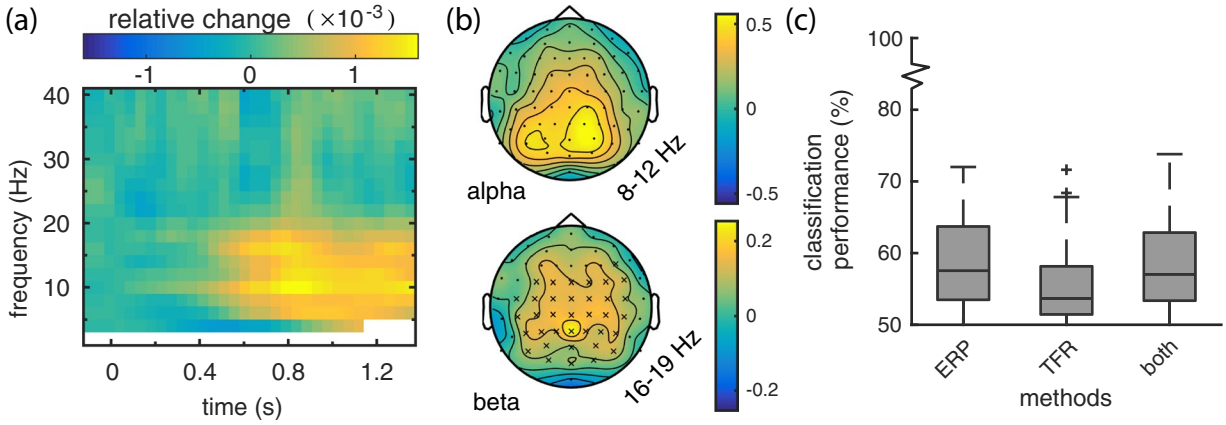

Figure 4. Results from the time-frequency analysis. (**a**) Grand average TFR of related probes subtracted from unrelated probes relative to the summed power per frequency band, averaged across electrodes (frequency dependent window length). (**b**) Topographies for the alpha (8–12 Hz) and beta (16–19 Hz) bands. Electrodes belonging to an identified cluster are marked by '×'. **(c)** Classification results of the ERP classifier, the TFR classifier, and a classifier trained on both feature sets.

rate at this level (only the condition level contrast)[23]. Therefore, to aid interpretation, and in particular, to be able to compare our results to another study reporting effects in the time-frequency domain[21], we performed two post-hoc significance tests, in which we isolated the alpha and beta frequency band respectively, and applied two cluster permutation tests, each on data averaged over the relevant frequency bands, and timepoints for the full trial (0 to 1.35 s). We Bonferroni corrected the significance values for performing these two tests, resulting in a significant effect for the beta band ($p = 0.001$), but with a non-significant cluster ($p = 0.015$) for the alpha band, at the two-tailed $\alpha/2 = 0.0125$ level (two-tailed tests, Bonferonni-corrected $\alpha = 0.025$, $n = 20$). These results are plotted in Fig. 4b, with electrodes in a significant cluster marked with '×'. In Fig. 4b, we see that the raw increase in power was larger for the alpha band, than for the beta (see the respective colorbar labels). However, the post-hoc significance tests determine that, though less pronounced, the difference in the beta band is, in fact, significantly different.

To determine if this difference in oscillatory behaviour can be decoded from single probe presentations, we performed a classification analysis on this TFR data, analogously to the ERP data classification, and also created classifiers for both types of data combined. The results of this analysis can be found in Fig. 4c. In the first boxplot we see a summary of the classification results shown earlier in Fig. 3a, together with the TFR results and classifier accuracies from the combined data. It appears to be possible to predict the relatedness of a probe based on the time-frequency data for certain participants, but overall results look worse than for the ERP based classifiers. Furthermore, it looks like there is no benefit from including this data in addition to the ERP data when using a classifier.

Accumulating predictions across probes in a trial. We have now generated predictions for single probes to estimate how well we can classify related from unrelated probes. However, for a BCI, information across multiple probes would need to be integrated to ultimately infer the original target. We can simulate such an analysis with our data, by using the consecutive predictions from a trial to predict which trial sequence they are from.

Each trial has a pattern of related and unrelated probes occurring in specific positions. We can try to predict from which trial a set of consecutive probe predictions was obtained, by comparing the similarity between the consecutive predictions and the true relatedness of all presented trials. That is, for each trial, we compute a similarity as the inner product between the probe relatedness scores predicted by the classifier (raw decision values), and the 'true' relatedness expected for every possible trial, given the sequence of probe words. We then rank the trials based on this similarity (high to low). We can use this rank as an estimate of how well trials can be predicted from the sequence of probe predictions, by interpreting it as a percentile: a trial that receives a high percentile rank (e.g., 99th rank), obtained (among) the highest score on the basis of its predictions.

To show how this estimate changes as trials contain more probes, we start by only considering the first probe of each trial, then both the first and second probe, and so on, until all 10 probes are included in this analysis. In Fig. 5(a), the results for the probe predictions on the individual participants' data can be found, together with the results for randomly generated probe predictions (1000 randomisations). Note that the maximum percentile rank that can be obtained is dependent on the total number of possible patterns. This does not necessarily correspond to the total number of trials, as trials may have duplicate patterns. This ceiling is plotted in the figure, together with the number of unique patterns per number of consecutive probes considered.

The figure shows that for random predictions it does not matter how many predictions are accumulated for the percentile rank of the correct trial. For actual predictions, however, the (mean) percentile rank in which the correct trial is found increases when more consecutive probes are considered. There are also participants (in the tail of the boxplots) for which, even with 10 consecutive probes, the percentile rank cannot be distinguished from the random prediction data.

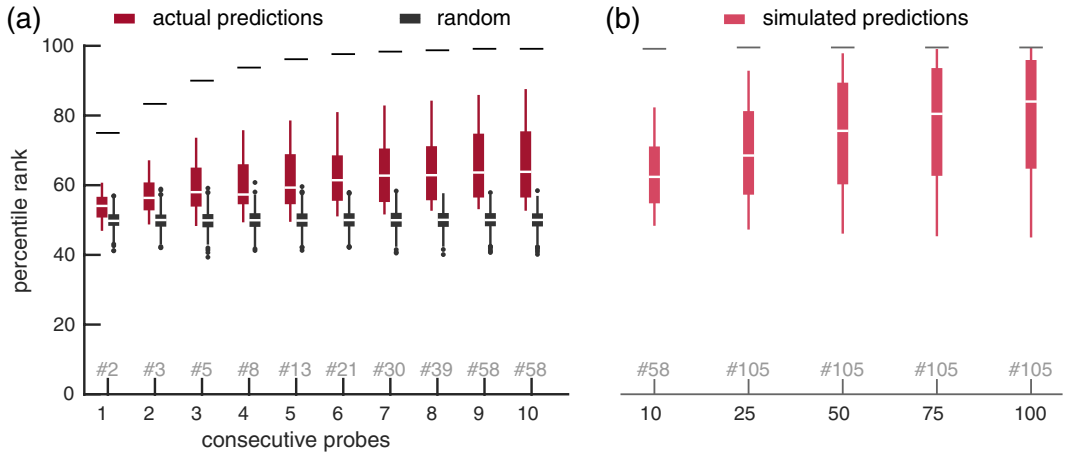

Figure 5. Percentile rank scores for predicting a trial, with varying number of consecutive probes considered. (**a**) Percentile ranks based on either the participants' probe predictions, or on randomly generated probe predictions are depicted. Horizontal lines denote the highest percentile rank that can be achieved given the number of unique trials. The amount of unique trials varies based on the number of consecutive probes that are considered, denoted in gray with a '#'. (**b**) Results from simulated data, projected beyond the 10 probes in the study, using simulated trial sequences of varying length (10, 25, 50, 75 and 100 probes). Predictions for these trials were generated for each participant, modelled by their classification accuracies.

In Fig. 5(b), we use simulated data to extend this analysis beyond 10 consecutive probes, to 25, 50, 75 and a 100 probes. This simulation is achieved by generating longer trials (permuted concatenations of existing trials), and using each participant's classification accuracy to simulate their responses to these hypothetical probes. As trials extend, the difference between participants grows. An average percentile rank of ∼97%, means that correct trial was on average within the top 3 predicted trial sequences (given a 105 unique trials). For 25, 50, 75 and a 100 probes, this was achieved for 0, 2, 3 and 5 subjects, respectively.

Discussion

In this study we aimed to determine whether the N400 can be elicited reliably using multiple probes after an initial target. Overall, our results show that there is a difference in brain response to related versus unrelated probes, as evidenced by both the grand average ERPs and the classification results. Specifically, in the grand average ERP we see an early component, more negative for unrelated stimuli, that matches the N400 in timing and location. Furthermore, the amplitude of this N400 effect is similar for probes appearing at the start and end of a trial (Fig. 2). In addition, classification results did not show any change with probe position within a trial. Together, these results allow us to conclude that the magnitude of the N400 does not diminish over sequences of probes even when another, potentially distracting, semantic context was presented in between.

This result is consistent with a recent publication[24], aiming to decode word relevance to a category of interest using sequential presentation of words and Fixation Related Potentials (i.e., ERPs time-locked to eye fixation on the stimulus). The authors found a more negative ERP for non-relevant words, similar to the N400. Interestingly, the results show markedly different brain responses for the online phase, where longer sequences were used (100 words), compared to the calibration phase (22 words). This may still be an indication that the N400 amplitude reduces for very long sequences (e.g., >10). Alternatively, with only a limited number of words used per category, stimuli were likely repeated frequently, which is known to cause a reduction in the N400 amplitude[16].

An unexpected result in our initial analysis was an overlapping late-negative component occurring after related probes (see Figs 1a and 2b). This component may be useful for BCI applications if it represents an additional signal to distinguish related from unrelated probes. Alternatively, as it appears most strongly for the final probes in a sequence, it may be an experimental confound reflecting an expectation of trial end. In future work we aim to investigate these possibilities further. In this work, to alleviate this late component overlapping with ERPs from the subsequent stimulus, we post-hoc excluded unrelated probes that appeared after a related probe. This reduced the number of unrelated probes in the final (53 probes) versus starting positions (104 probes).

Previous research has indicated that N400-tasks may also induce changes in power in the frequency domain[21]. When decomposing our data into the frequency domain, we visually observe *higher* power for related probes in the alpha and beta band from 600 to 1200 ms post stimulus presentation, though, only the beta band (16–19 Hz) increase was statistically significant. This is in contrast to Wang *et al.*[21] who showed a *decrease* in beta power for in-congruent sentence endings (that elicit a more negative N400). This inconsistency in response suggests that either the role of beta power in relation to the N400 is different for sentence close-word paradigms, or that another aspect of our task (e.g., behavioural response preparation) results in these changes in beta power. These conflicting results, together with the classification analysis which showed no additional benefit from the inclusion of time-frequency data over ERP data only, suggest only limited usability of this response for a BCI.

Agreement between the behavioural responses and the probe labels ranged from 70% to 95% across participants (mean 87%). Disagreement can reflect an error, or an inherent disagreement on the relatedness of the most

recent probe and the target. Participants with higher percentage agreement, thus, may have been better at keeping the target on mind, or have more conventional relatedness judgements. The significant correlation between classification accuracy and behavioural agreement ($r = 0.58$), can in fact be explained by both interpretations.

The classification results overall, show relatively low accuracies ($\mu = 58\%$) and a large variability across subjects (50–72%). The fact that for about a third of subjects their accuracy could not significantly be distinguished from chance is a concern for using any BCI application on the basis of this paradigm. There is some precedent in difficulty to detect the N400 within individual subjects in other N400 experimental paradigms[25], and having a subset of users for whom the intended brain signal cannot be detected well, is a not an uncommon problem for BCIs (see e.g.[26] for a discussion).

One approach to overcome low accuracies in a BCI is to aggregate results over multiple stimulus repetitions. However, as the N400 amplitude can be reduced by stimulus repetition[16], we instead accumulate over a sequence of different probes, as shown in the analysis in Fig. 5. By generating longer trials and simulating classifier predictions for each subject, we then extrapolated the results beyond the 10 probes used in this study. These results (Fig. 5b), suggest that accumulation can be effective but *only* for subjects with sufficiently high single trial accuracy. Assuming these simulations are valid, this implies this type of semantic BCI would be usable (true target in the top-3 in <=100 probes) for only a small subset of participants (2–5 out of 20). In reality such long sequences present additional challenges, such as subject motivation and attention decrease over the trial, which may reduce performance further. On the other hand, a closed-loop BCI, could use the information accumulated from previous probes to select (the most) informative future probes, which may reduce the required sequence length compared to the non-optimised probe sequences used in these simulations. This remains an area for future research.

While practical utility as a BCI that infers the target word may be limited, the consecutive probing paradigm may have alternate applications. For instance, in assessing linguistic processing in patients with Disorders of Consciousness[27,28]. For this purpose, word-pair priming or sentence congruence paradigms have previously been used, and the presence of the N400 in these tasks has been shown to correlate with recovery[28]. The sequential probing paradigm used in our experiment, might be able to detect higher level cognitive processing in patients that exhibit a N400, by determining whether or not these patients can hold a target active over multiple consecutive probe words. These approaches may similarly be useful for assessment of patients in a Complete Locked in State[29].

In summary, our results show that we can decode responses of relatedness from EEG, on the basis of the N400, and we find no indication that this response attenuates across multiple consecutive probes. While this paradigm can in theory be adapted to a BCI that infers the target word on a user's mind, low singe trial accuracies and high variability in accuracy across subjects make this unlikely to offer practical utility in communication applications. However, the paradigm itself may yet be of use for assessing cognitive processing in certain patients populations.

Methods

Participants. Twenty-one participants took part in the experiment (12 female, 9 male), ranging in age from 18 to 61 years old (mean age 29, sd = 12). Participants were only eligible to participate if they were native Dutch speakers and reported to have no reading problems (e.g., dyslexia)

One participant dropped out of the experiment due to excessive sleepiness. Her data are not included in the analysis. Response data are missing for participant 9, block 6, due to a technical problem, and there is no data for participant 19, block 6 (due to initial cap fitting difficulties, the experiment could not be completed in the allotted time-frame).

All participants provided informed consent prior to participation, and the study was approved by and conducted in accordance with the guidelines of the Ethical Committee of the Faculty of Social Sciences at the Radboud University.

Task. In the experiment participants completed 212 trials, each consisting of a target word, followed by one to ten probe words. For each trial, participants were tasked to remember the target word, and subsequently mentally assess for each probe whether or not it was semantically related to the target. At the end of a trial participants were prompted to specify by button-press the relatedness status of the most recent probe (left = related, right = unrelated). 50% of trials contained ten probes, while the remaining 50% contained an equal distribution of one to ten probes. Trials with fewer probes were included to ensure that participants would mentally evaluate all probes, and not only those for which they could predict that a question would follow.

Participants were instructed to respond as fast as possible when prompted with a question, and received their reaction time as feedback (this feedback was coloured using a gradient: green to white to red at 700 ms, 950 ms, 1200 ms, respectively). This feedback did not reflect 'correctness' of the choice as relatedness judgements can vary between individuals. To increase salience and facilitate memorisation, participants were asked to pronounce the target word during its presentation.

Each target word was presented for 2000 ms, followed by a 1000 ms interstimulus interval (ISI). Each subsequent probe was presented for 350–370 ms (jittered duration) and also followed by a 1000 ms ISI. The question prompt was displayed after the last ISI, and was replaced by feedback on buttonpress. Across the 212 trials a total of 539 related, and 1072 unrelated probes were presented (33.5% related).

In our analyses we compare probes occurring in the first position with probes occurring in the 'final' position (i.e., 9^{th} and 10^{th} position). Note that we define the 'final' category to consist of two probe positions, in order to get an equal number of brain responses, as only 50% of trials consisted of the full 10 probes. With this analysis in mind we designed trials such that there was an equal distribution of related and unrelated probes for the first and final position(s). Finally, these trials were set up such that no two related probes were ever presented in a row, as a brain response signifying relatedness in such a case could be purely from the pairing with the previous stimulus. The percentage of related probes per probe position can be found in Fig. 6a.

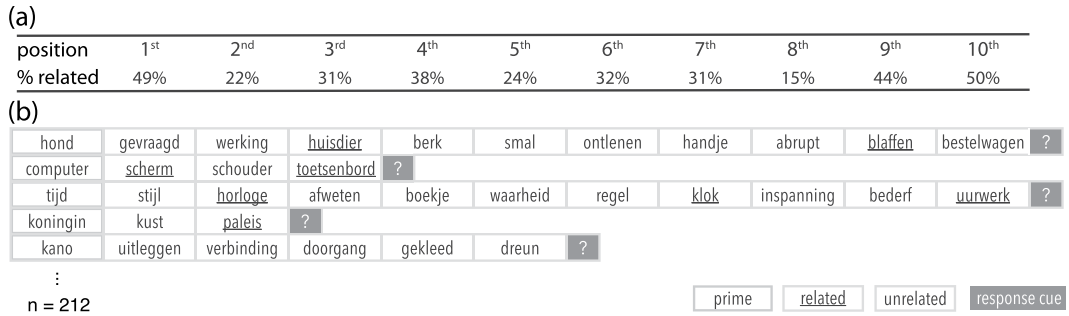

Figure 6. Trial design with example stimuli (**a**) Percentage of related stimuli per probe position in the trial. (**b**) Example trials: A trial consists of an initial target, followed by up to 10 probes. A probe is either related or unrelated to the target. Each trial ends with a behavioural prompt. A translation of the first trial in english: '**dog**' - 'asked', 'operation', 'pet', 'birch tree', 'narrow', 'derive', 'hand', 'abrupt', 'bark', 'delivery van'.

Stimulus words were obtained from the Leuven association dataset[30]. This Dutch dataset consists of cue-words and responses from people who were asked to specify (up to) three words that a given cue-word brought to mind. In our design we used Leuven cue-words as targets, and Leuven response-words as (related) probes. We then used the CELEX Dutch Wordform database[31] to obtain words unrelated the the targets, to use as unrelated probes. Probes never occurred more than once across all trials, as repetition of concepts has been shown to affect the N400 amplitude[16], though some targets also occurred as probes. Unrelated words were selected to approximate the related probes in distribution of both word length and word frequency. Word frequency in particular, is known to affect N400 amplitudes, with words that occur frequently in a language eliciting smaller N400s[14,15]. Across all probes, the mean (log) word-frequency was 1.50 (sd = 0.65) for related, and 1.49 (sd = 0.63) for unrelated probes. On average the length of a related probe was 5.68 characters (sd = 1.98), and an unrelated probe 6.13 characters (sd = 1.92).

Stimuli for a few example trials are depicted in Fig. 6b. A detailed description of the selection procedure and a full list of stimuli can be found in the Supplementary Information.

Trial order was randomised per subject.

Equipment. EEG was recorded with 64 sintered Ag/AgCl active electrodes (BioSemi, Amsterdam, The Netherlands), placed according to the international 10–20 system, at a sampling rate of 512 Hz. The Biosemi ActiveTwo system uses a Common Mode Sense (CMS) and a Driven Right Leg (DRL) electrode, instead of a ground electrode. The recorded signals reflect the voltages between each electrode and the CMS, which can subsequently be re-referenced to any electrode(s) of choice. We placed two electrodes on the mastoids for this purpose and used four further electrodes to measure vertical and horizontal EOG.

Data Analysis. *Preprocessing.* Data were recorded in 6 blocks of 35 trial sequences each. These data blocks were loaded, highpass filtered at 0.1 Hz (4th order Butterworth filter)[32], and then sliced into segments (i.e., epochs), from 1.5 s prior to and 2 s after each probe presentation. Then, data from different blocks were combined and downsampled to 256 Hz. All electrodes were mean-corrected (centered) and re-referenced to the average of the mastoid electrodes. The four EOG channels were then regressed-out of the EEG channels to remove any influence from eye movements or blinks[33]. The EEG channels were subsequently lowpass filtered at 40 Hz (6th order Butterworth filter).

After these preprocessing steps, any epochs or channels with abnormal activity were marked for removal (epochs) or interpolation (channels). A trial or channel was considered abnormal when the variance of the given unit diverged more than 3.5 standard deviations from the median of all units (channels or epochs). This was repeated, excluding previously marked units in the next iteration, until no units were considered abnormal (max 6 iterations). This resulted in between 12 and 156 epochs and up to 4 channels to be marked per participant. Any identified bad channels were replaced with an interpolated'virtual channel' using a spherical spline interpolation[34], while bad epochs were marked to be ignored during grand average calculation and training of classifiers. Finally, epochs were baselined to a period from −100 ms to 0 ms from stimulus onset.

Grand Averages. For brain responses in the time domain, we obtained grand average ERPs by averaging across relevant probes, and then averaging across subjects. The data for the grand average ERPs were low-pass filtered at 20 Hz, prior to averaging, for plotting purposes (smoothing). To analyse brain responses in the frequency domain, the data from each trial was decomposed into frequency bins of 2 Hz, from 2 to 40 Hz. The power in these bins was calculated for each 50 ms, starting from 0.5 s prior to probe onset to 1.35 s after, using a Hanning window and a frequency dependent window length (ranging 3.5 s - to 175 ms respectively). These Time Frequency Representations (TRFs) were then averaged across participants to obtain grand average TRFs.

Significance testing. To determine whether brain responses between conditions (e.g., related vs unrelated) were significantly different, we used non-parametric cluster-based permutation tests[22] (as implemented in Fieldtrip[35]). Such a test allows for the combination of information across electrodes and timepoints, to increase sensitivity of the statistical test, without having to correct for multiple comparisons with respect to those aspects. These tests

were performed using a within-subject design, and a dependent samples t-test as test-statistic. Note that this non-parametric cluster-based permutation test does not rely on assumptions with regard to the distribution of the data, regardless of the chosen test-statistic (the test-statistic is merely used to quantify the difference between datapoints). In all cases, we performed a two-tailed test, (tail-corrected alpha = 0.025), using 1000 permutations, supplying all channels, and timepoints and/or frequency bins as specified.

Classification. For classification analyses on data in the time domain, the preprocessed data were further downsampled to 64 Hz and low-pass filtered to 15 Hz to reduce the amount and complexity of data passed to the classifier and prevent overfitting. We use a classification pipeline identified as robust for different types of ERPs[36], consisting of a spectral filter, a spatial whitening and classification using a regularized classifier. Specifically, the preprocessed data from time 0 to 1 s and all remaining epochs and channels (already spectrally filtered), were spatially whitened, and subsequently classified using a L2-regularized logistic regression. The regularization parameter was optimised using a 5-fold (nested) crossvalidation. A separate classifier was trained for each participant. For classification analysis on data in the time-frequency domain, preprocessed data were again decomposed into frequency bins of 2 Hz (2–40 Hz), but here the bins were calculated for each 100 ms from probe onset (0 s), until the end of the trial (1.35 s), using a Hanning window and a frequency dependent window length (1 s–175 ms respectively). These time-frequency data were then processed using the same classification pipeline.

To estimate classification performance, the data were separated into crossvalidation folds of training and test sets, for each participant, where all but one trial sequences were used as training data, and the data from probes in the excluded trial sequence were used as the test set. Only trial sequences in which ten probes were presented were used as test sets, to avoid biasing the results toward responses to early probes. Epochs that were marked for removal during preprocessing were excluded when occurring in a training set, but were included when part of the test set (in an online BCI application poor data quality does not always preclude an accurate prediction).

Data Availability

The dataset and accompanying analysis files generated during the current study are available in the *Donders Repository*: http://hdl.handle.net/11633/di.dcc.DSC_2016.00314_172.

References

1. Miner, L. A., McFarland, D. J. & Wolpaw, J. R. Answering questions with an electroencephalogram-based brain-computer interface. *Archives of Physical Medicine and Rehabilitation* **79**, 1029–1033, https://doi.org/10.1016/S0003-9993(98)90165-4 (1998).
2. Hill, N. J. *et al*. A practical, intuitive brain–computer interface for communicating 'yes' or 'no' by listening. *Journal of Neural Engineering* **11**, 035003, https://doi.org/10.1088/1741-2560/11/3/035003 (2014).
3. Chaudhary, U., Xia, B., Silvoni, S., Cohen, L. G. & Birbaumer, N. Brain–Computer Interface–Based Communication in the Completely Locked-In State. *PLOS Biology* **15**, e1002593, https://doi.org/10.1371/journal.pbio.1002593 (2017).
4. Farwell, L. A. & Donchin, E. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology* **70**, 510–523 (1988).
5. Thielen, J., Broek, Pvd, Farquhar, J. & Desain, P. Broad-Band Visually Evoked Potentials: Re(con)volution in Brain-Computer Interfacing. *PLOS ONE* **10**, e0133797, https://doi.org/10.1371/journal.pone.0133797 (2015).
6. Waal, Mvd, Severens, M., Geuze, J. & Desain, P. Introducing the tactile speller: an ERP-based brain–computer interface for communication. *Journal of Neural Engineering* **9**, 045002, https://doi.org/10.1088/1741-2560/9/4/045002 (2012).
7. Acqualagna, L. & Blankertz, B. A gaze independent spelling based on rapid serial visual presentation. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC*, 4560–4563, https://doi.org/10.1109/IEMBS.2011.6091129 (2011).
8. Laine, M. & Martin, N. *Anomia: Theoretical and Clinical Aspects* (Psychology Press, 2013).
9. Geuze, J., Farquhar, J. & Desain, P. Towards a Communication Brain Computer Interface Based on Semantic Relations. *PLoS ONE* **9**, e87511, https://doi.org/10.1371/journal.pone.0087511 (2014).
10. Kutas, M. & Federmeier, K. D. Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology* **62**, 621–647, https://doi.org/10.1146/annurev.psych.093008.131123 (2011).
11. Ganis, G., Kutas, M. & Sereno, M. I. The Search for "Common Sense": An Electrophysiological Study of the Comprehension of Words and Pictures in Reading. *Journal of Cognitive Neuroscience* **8**, 89–106, https://doi.org/10.1162/jocn.1996.8.2.89 (1996).
12. Holcomb, P. J. & Neville, H. J. Auditory and Visual Semantic Priming in Lexical Decision: A Comparison Using Event-related Brain Potentials. *Language and Cognitive Processes* **5**, 281–312, https://doi.org/10.1080/01690969008407065 (1990).
13. Hagoort, P. & Brown, C. M. ERP effects of listening to speech: semantic ERP effects. *Neuropsychologia* **38**, 1518–1530, https://doi.org/10.1016/S0028-3932(00)00052-X (2000).
14. Van Petten, C. & Kutas, M. Interactions between sentence context and word frequencyinevent-related brainpotentials. *Memory & Cognition* **18**, 380–393, https://doi.org/10.3758/BF03197127 (1990).
15. Laszlo, S. & Federmeier, K. D. Never seem to find the time: evaluating the physiological time course of visual word recognition with regression analysis of single-item event-related potentials. *Language, Cognition and Neuroscience* **29**, 642–661, https://doi.org/10.1080/01690965.2013.866259 (2014).
16. Rugg, M. D. Event-related brain potentials dissociate repetition effects of high-and low-frequency words. *Memory & Cognition* **18**, 367–379, https://doi.org/10.3758/BF03197126 (1990).
17. Deacon, D., Breton, F., Ritter, W. & Vaughan, H. G. The Relationship Between N2 and N400: Scalp Distribution, Stimulus Probability, and Task Relevance. *Psychophysiology* **28**, 185–200, https://doi.org/10.1111/j.1469-8986.1991.tb00411.x (1991).
18. Deacon, D. & Shelley-Tremblay, J. How automatically is meaning accessed: a review of the effects of attention on semantic processing. *Frontiers in Bioscience* **5**, 82–94 (2000).
19. van Vliet, M., Mühl, C., Reuderink, B. & Poel, M. Guessing What's on Your Mind: Using the N400 in Brain Computer Interfaces. In Yao, Y. *et al*. (eds.) *Brain Informatics*, Lecture Notes in Computer Science, 180–191 (Springer Berlin Heidelberg, 2010).
20. Geuze, J., van Gerven, M. A. J., Farquhar, J. & Desain, P. Detecting Semantic Priming at the Single-Trial Level. *PLoS ONE* **8**, e60377, https://doi.org/10.1371/journal.pone.0060377 (2013).
21. Wang, L. *et al*. Beta oscillations relate to the N400m during language comprehension. *Human Brain Mapping* **33**, 2898–2912, https://doi.org/10.1002/hbm.21410 (2012).
22. Maris, E. & Oostenveld, R. Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods* **164**, 177–190, https://doi.org/10.1016/j.jneumeth.2007.03.024 (2007).
23. Maris, E. Statistical testing in electrophysiological studies. *Psychophysiology* **49**, 549–565, https://doi.org/10.1111/j.1469-8986.2011.01320.x (2012).

24. Wenzel, M. A., Bogojeski, M. & Blankertz, B. Real-time inference of word relevance from electroencephalogram and eye gaze. *Journal of Neural Engineering* https://doi.org/10.1088/1741-2552/aa7590 (2017).
25. Cruse, D. *et al*. The reliability of the N400 in single subjects: Implications for patients with disorders of consciousness. *NeuroImage: Clinical* **4**, 788–799, https://doi.org/10.1016/j.nicl.2014.05.001 (2014).
26. Allison, B. Z. & Neuper, C. Could Anyone Use a BCI? In *Brain-Computer Interfaces*, Human-Computer Interaction Series, 35–54 (Springer, London, 2010). https://doi.org/10.1007/978-1-84996-272-8_3.
27. Kotchoubey, B. *et al*. Information processing in severe disorders of consciousness: Vegetative state and minimally conscious state. *Clinical Neurophysiology* **116**, 2441–2453, https://doi.org/10.1016/j.clinph.2005.03.028 (2005).
28. Steppacher, I. *et al*. N400 predicts recovery from disorders of consciousness. *Annals of Neurology* **73**, 594–602, https://doi.org/10.1002/ana.23835 (2013).
29. Kübler, A. & Birbaumer, N. Brain–computer interfaces and communication in paralysis: Extinction of goal directed thinking in completely paralysed patients? *Clinical Neurophysiology* **119**, 2658–2666, https://doi.org/10.1016/j.clinph.2008.06.019 (2008).
30. Deyne, S. D. & Storms, G. Word associations: Norms for 1,424 Dutch words in a continuous task. *Behavior Research Methods* **40**, 198–205, https://doi.org/10.3758/BRM.40.1.198 (2008).
31. Baayen, R. H., Piepenbrock, R. & Gulikers, L. The celex lexical database (release 2). *Distributed by the Linguistic Data Consortium, University of Pennsylvania* (1995).
32. Tanner, D., Morgan-Short, K. & Luck, S. J. How inappropriate high-pass filters can produce artifactual effects and incorrect conclusions in ERP studies of language and cognition. *Psychophysiology* **52**, 997–1009, https://doi.org/10.1111/psyp.12437 (2015).
33. Gratton, G. Dealing with artifacts: The EOG contamination of the event-related brain potential. *Behavior Research Methods, Instruments, & Computers* **30**, 44–53, https://doi.org/10.3758/BF03209415 (1998).
34. Perrin, F., Pernier, J., Bertrand, O. & Echallier, J. F. Spherical splines for scalp potential and current density mapping. *Electroencephalography and Clinical Neurophysiology* **72**, 184–187, https://doi.org/10.1016/0013-4694(89)90180-6 (1989).
35. Oostenveld, R., Fries, P., Maris, E. & Schoffelen, J.-M. FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data (2011). https://doi.org/10.1155/2011/156869.
36. Farquhar, J. & Hill, N. J. Interactions Between Pre-Processing and Classification Methods for Event-Related-Potential Classification. *Neuroinformatics* **11**, 175–192, https://doi.org/10.1007/s12021-012-9171-0 (2013).

Acknowledgements

We would like to thank James McQueen for valuable feedback on an earlier draft of this paper.

Author Contributions

K.D., J.F. and P.D. conceived the experiment, K.D. conducted the experiment, analysed the results and wrote the manuscript. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-51011-4.

Competing Interests: The authors declare no competing interests.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.