Brainprint: Identifying Unique Features of Neural Activity with Machine Learning

Maria Ruiz-Blondet (mruizbl1@binghamton.edu)[1]
Negin Khalifian (nkhalif1@binghamton.edu)[2]
Blair C. Armstrong (b.armstrong@bcbl.edu)[3]
Zhanpeng Jin (zjin@binghamton.edu)[1, 4]
Kenneth J. Kurtz (kkurtz@binghamton.edu)[2]
Sarah Laszlo (slaszlo@binghamton.edu)[2]

[1]Department of Bioengineering, 4400 Vestal Parkway East, Binghamton, NY 13902 USA
[2]Department of Psychology, 4400 Vestal Parkway East, Binghamton, NY 13902 USA
[3]Basque Center on Cognition, Brain, and Language, Paseo Mikeletegi 69, 2nd Floor, 20009 DONOSTIA SPAIN
[4]Department of Electrical and Computer Engineering, 4400 Vestal Parkway East, Binghamton, NY 13902 USA

Abstract

Can a person be identified uniquely by some feature of their neural activity, as they can be by fingerprints? If so, 1) what would those features be like and 2) are existing computational methods sufficient to extract them? Here, we explore these questions by coordinating psychophysiological and machine learning approaches. We begin with the proposition that one unique feature of individual cognition is the detailed network of concepts, and relationships between concepts, that are present in each individual's semantic memory. We then demonstrate that we are able to accurately classify individual unlabeled brain activity—in the form of Event-Related Potentials (ERPs) elicited during a task that probes semantic memory—to the individual it belongs to with several pattern classifiers. These results demonstrate that it is possible to identify individuals on the basis of unique features of their brain activity. Biometric applications are discussed.

Keywords: Machine Learning; Event-Related Potentials; Individual Differences; Biometrics

Introduction

Each of us has a sense that our individual cognitive worlds--our selves-- are unique. In a materialist epistemological framework, the self is instantiated by the brain, and thus it is the brain and its workings that make our selves unique. This sense is codified in modern cognitive science and cognitive neuroscience by the idea of individual differences, which is the acknowledgement that not all brains are identical, and not all individuals engage cognitive processes in an identical manner (see, for example, Daneman & Carpenter, 1980; Raz et al., 2005). The idea of individual differences is taken to its extreme in the field of biometrics, where it is assumed not only that individuals differ on some measure (e.g., fingerprint or retinal topography), but that individuals are unique on those measures. In this field, the EEG signal is starting to be considered as an identification characteristic (see for example, Jian-feng, 2010; Palaniappan & Mandic, 2007); however, the prior work has not substantially interfaced with what is understood about the cognitive processes that impact an individual's EEG or ERPs.

One well-understood cognitive system that seems likely to differ uniquely between individuals is semantic memory, defined here as a memory network of concepts and relationships between concepts. As an example of how individuals' semantic networks differ, consider the concepts [bee] and [anaphalaxis]. Even with only these two concepts, there are a number of plausible associated states in semantic memory that might be instantiated in individuals: an individual might not know what either of these things are, or might know both of them, or only one or the other. A person with a bee allergy might strongly associate the two concepts, while a person with no bee allergy might not associate them at all, or associate them only weakly. Of course, there are many more concepts and relationships that can be represented than bees and anaphalaxis, and the more concepts and relationships that need to be represented, the less likely any two people are to represent them in exactly the same way (i.e., as the pool of possible concepts grows beyond only [bee, anaphalaxis] to a larger pool like [bee, anaphalaxis, snake, spider, chocolate, prawn, cilantro, clown], and then to larger pools and so on, it becomes less and less likely for all concepts and relationships to be represented the same way in multiple individuals). Although it seems plausible that there are numerous neuro-cognitive systems that distinguish between individuals, semantic memory, as we will see below, produces a large, robustly studied electrophysiological response that has already been demonstrated to vary across individuals (although not necessarily produce unique responses across individuals).

A Neural Measure of Semantic Memory

Attempts to access semantic memory are known to elicit a large, robust, electrophysiological response known as the N400. The N400 is a negative going Event-Related Potential (ERP) component that peaks at approximately 400 ms post stimulus onset, and is maximal over the back of the head. The N400 is strongly sensitive to numerous manipulations of semantic memory including, but not limited to, violations of sentence context, semantic priming, imageability, concreteness, and number and strength of lexical associations (see Kutas & Federmeier, 2011, for review). Importantly, we have demonstrated in past work that the N400 is sensitive to whether or not a particular visual word form has meaning to an individual participant. Specifically, when participants are presented with a large variety of acronyms (e.g., DVD, NPR), individuals present larger N400s to acronyms they are familiar with than to acronyms they are not familiar with (see Laszlo & Federmeier, 2007). This is taken to mean that when an unfamiliar item is presented, the system is less able to make contact with concepts in semantic memory than when a familiar item is presented.

Participants are able to identify, on average, 83% of 75 acronym items in our stimulus list-- 62 items. There are 1.2 x 1014 possible ways to randomly choose 62 items from a set of 75, which quantifies the idea that it is very unlikely for any 2 people to have an identical profile of familiar and unfamiliar items when the set of items is of sufficient size (and, in fact, no two participants had exactly the same pattern of known and unknown items in this dataset). In what follows, we exploit the known variation in what acronyms are familiar to individual participants as one possible source of a signal that is unique to individuals. Other sources include individual variation in neural anatomy that result in slightly different sizes and distributions of ERPs across the scalp and slightly different timing of the N400 and the ERP components that precede it -- each of these factors is represented in the data on which we performed pattern classification.

Machine Learning: Pattern Classification

A wide variety of pattern classifying algorithms exist that could, in principle, be applied to the problem under study (for extensive review, see Bishop, 1995). Here, we focus on the performance of three methods that past work suggests should be well-suited for identifying unique features in distributed, high-dimensional representations of neural activity (such as the temporally extended ERP signal). The simplest method we considered was creating a simple linear discriminant based on the normalized cross-correlation between pairs of waveforms (i.e., labeling a test waveform as belonging to an individual if that waveform had a higher cross-correlation with another waveform from the same individual than with waveforms from any other individuals). This method is based on the intuitive notion that, if brainprinting is possible, overall, waveforms elicited by the same person should be more similar than waveforms elicited by different people, and also on past work suggesting that cross-correlation is an effective means of measuring EEG waveform similarity (e.g., Chandaka & Chatterjee, 2009). However, this method is not especially flexible; for example, it gives equal weight to similarities in all portions of the waveform, even though the most important similarities-- those reflecting similar semantic memory networks-- should occur in temporally specific portions of the waveform, and should therefore likely be weighted more heavily by the pattern classifier.

Pattern classifiers with increased flexibility, such as the ability to learn, are therefore appealing for the brainprint problem. It is advantageous for the classifier to be able to learn what parts of the waveform are most important in telling people apart, and what parts are either not informative or anti-informative. Here, we considered two learning classifiers that seemed to us to be particularly likely to be able to solve this problem. These are Divergent Autoencoder (DIVA; Kurtz, 2007) and Naive Discriminant Learning (see Baayen, Milin, Durdevic, Hendrix, & Marelli, 2011). The divergent autoencoder is a feed-forward neural network architecture that provides an alternative to the multilayer perceptron (MLP) for applying the backpropagation algorithm to classification tasks. In contrast to the MLP, which has a single output node for each possible classification, DIVA has a full copy of the input layer (a “channel”) corresponding to each possible classification. The key design principle is training the autoencoder to reconstruct the members of each category with the constraint that each autoassociative channel shares a common hidden layer. Classification outcomes are a function of reconstruction error: an item is a good member of a class to the extent it can be recoded and decoded along the appropriate channel with minimal distortion. Kurtz (2007) originally developed DIVA as a cognitive model of human category learning; research is currently in progress that establishes the wider potential of DIVA networks as a highly effective, general-purpose classifier for machine learning.

NDL was selected as an alternate method as it has recently received considerable attention both because of its ability to account for classification phenomena across domains and because its computational characteristics make it well-suited for modeling large

data sets that would typically be extremely computationally expensive for related connectionist models. This advantage is due in part to the derivation of equilibrium equations (Danks, 2003) that allow for the rapid calculation optimal weights, which enables the training of NDL models on extremely large data sets (because the input to the models here is an entire ERP waveform, for an input layer size of 550 units, the present problem is substantially larger in size than many cognitive modeling problems). Similar to many other machine learning algorithms, however, it is capable of learning to weight the contributions of different input dimensions based on their informativeness, allowing the algorithm to "focus in" on the most relevant dimensions of inputs for discrimination.

In what follows, we assess multiple metrics of accuracy for each of these classification methods in identifying unlabeled exemplar ERPs, with the goal of determining whether any of these techniques is able to learn to extract unique features of individual brain waves.

Method: ERPs

ERPs were drawn from an existing corpus of ERP visual word recognition data. These data were acquired in an experiment following the methods of our past studies demonstrating individual differences in N400s on the basis of individual acronym knowledge (Laszlo & Federmeier, 2007). In this study, EEG was recorded from 32 adult participants (11 female, age range 18-25, mean age 19.12) who silently read an unconnected list of text. EEG was digitized at 6 midline electrodes sites. Participants viewed 75 acronyms that each repeated once at a lag of 0, 2, or 3 intervening items, in addition to several other item types (words, pseudowords, and illegal consonsant strings) not analyzed in the present work. Participants were instructed to press a button on a gamepad when their name was presented on the screen. This task was given in order to ensure that participants were actively engaged in the experiment and attending to critical items (words, pseudowords, acronyms, and illegal strings) without contaminating waveforms elicited by critical items with response potentials.

That repetition was included in this design allows for homogenous but non-overlapping segmentation of the data into train and test corpora for machine learning: first responses to acronyms were used for training, and second responses were used for testing. ERPs were computed at each electrode time-locked to the onset of each of the four critical stimulus types, on each of the two presentations (e.g., words, first presentation; acronyms, second presentation). For a more detailed description of the methods, see Khalifian (2013).

Note that this experiment was designed primarily as a study of written language comprehension, *not* as a study of individual differences in psychophysiology. While the dataset does include responses that are likely to be non-identical across participants, as discussed, these differences were not maximized by design. For example, while the items are not likely to be represented identically by any two participants, they are relatively benign (e.g., DVD, NFL) and therefore not likely to elicit individual reactions that are particularly strong or idiosyncratic. A more targeted design might feature items more likely to have stronger individual differences; for example, words with strong affective loadings (e.g., SPIDER, CLOWN), or low frequency words likely to be known by some but not all participants (see Ramscar et al., 2014).

Similarly, because this experiment was not designed as a study of individual differences, relatively few trials were acquired from each participant, in anticipation of data analysis of group, as opposed to individual, averages, as is typical for ERP language experiments. The high signal to noise ratio in ERPs could prohibit meaningful averages from being formed from individuals with so few trials available. The non-targeted design of the corpus from which data for classification were drawn, as well as the relatively low signal to noise present in ERPs with so few trials, will both provide challenges to our classifiers. If we are able to achieve accurate classification in spite of these challenges, we will have reason to believe that our classification methods are robust.

Method: Pattern Classifiers

Training data for the classifiers was comprised of responses to the first presentation of acronyms; test data was comprised of responses to the second presentation of acronyms. The training and test data were thereby completely non-overlapping. After EEG artifact rejection, each participant contributed 70 trials to both data sets (some participants had more than 70 trials left after artifact rejection; from these 70 random trials were selected). One average per participant was not considered to be sufficient training data for the neural network classifiers. Therefore, 100 ERPs consisting of random averages of 50 of the 70 trials were made for each participant, resulting in 3200 averages (100 averages for each of the 32 participants) of 50 trials each to use as inputs for neural network training. Similarly, 100 random averages of 50 trials per participant were formed from the test data for network evaluation.

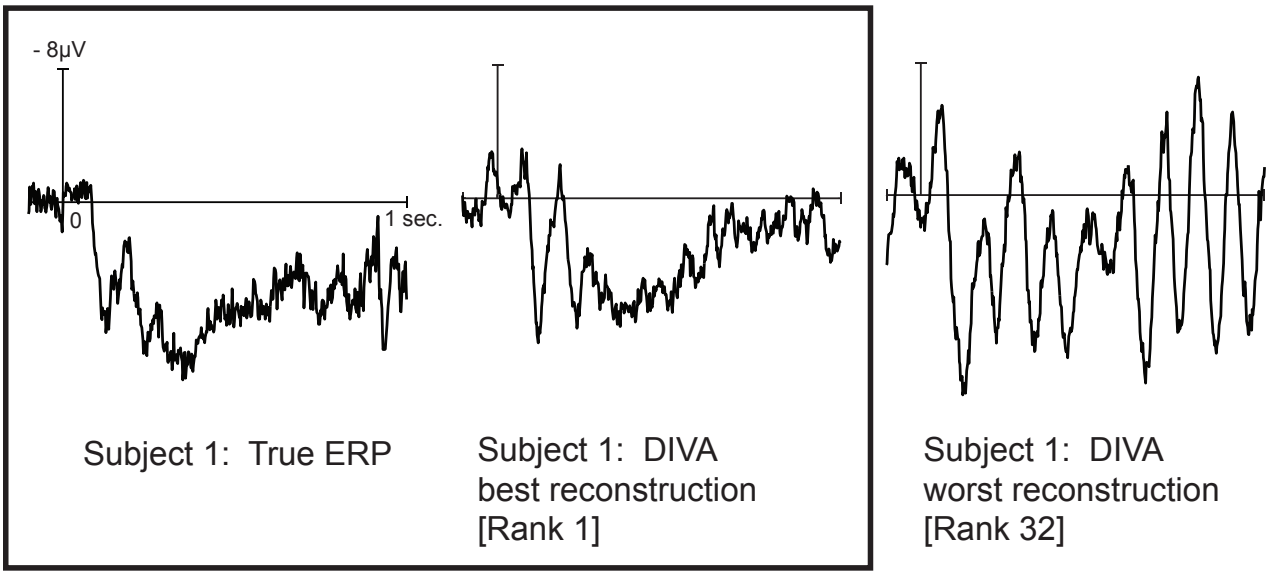

Figure 1. Sample Data and DIVA Reconstructions. On the left, a true ERP elicited by Subject 1. In the middle, the best DIVA reconstruction of that ERP after training. On the right, the worst DIVA reconstruction of that ERP after training. Notice that DIVA's best reconstruction appears as a slightly filtered version of the true ERP, with activity in early temporal epochs emphasized.

Cross-Correlation

To classify by cross-correlation, we first computed the maximum absolute value of the cross-correlation between pairs of waveforms (see Chandaka, Chatterjee, & Munshi, 2009). These pairs could be self-self pairs (i.e., one of the 100 averages from subject 1's training corpus and one of the 100 averages from subject 1's test corpus) or self-other pairs (i.e., one waveform elicited by subject 1 and another elicited by subject 2, or subject 3, and so on, for a total of 31 self-other pairs). The cross-correlations between pairs were then divided by the norm (or vector length) of the pair, in order to reduce variability caused by scalp thickness and other cognitive-unrelated events, allowing consistency in magnitude within cross-correlation results; data were also high-pass filtered during recording to eliminate variability due to DC shifts.. The output of this operation was then ranked, with the highest ranked pair being this classifier's guess as to which two waveforms were elicited by the same subject. This ranking method allows for the accuracy of the cross-correlation classifier to be analyzed identically to the accuracies of the other classifiers, as described below.

Divergent Autoencoding (DIVA)

The DIVA network was a 550:200:550[32] feedforward autoencoder. The 550 input nodes corresponded to the 550 samples in the ERP waveforms; thus the entire waveform was veridically presented to the network. The [32] signifies that, instead of having only one output layer representing the reconstruction of the input, as in a standard autoencoder, there were 32 output layers, one for each possible "classification" by the network of the input data (i.e., the model is making a 32-way classification; see Kurtz, 2007, for details). The 200 unit hidden layer was shown to be the smallest size that would enable near-perfect learning of the training set in pilot simulations. On each supervised training trial, hidden-to-output weights were adjusted only along the correct category channel. The input-to-hidden connections followed a sigmoidal activation function; the hidden-to-output connections followed a linear activation function. The network was trained for 1000 iterations; this was determined to be a level that allowed satisfactory (>99%) train performance without overfitting via prior validation simulations. After these 1000 iterations, weights in the model were fixed.

At test, the model was presented with each of the 3200 test examples, and activation was allowed to propagate through the network. Reconstruction error was measured on each output channel. The channel with the least output error was assigned rank 1 for that trial, the channel with the 2nd least output error was assigned rank 2 for that trial, and so on. Again, assigning ranks to the model's outputs allows for its accuracy to be analyzed in a manner identical to that used for the other classification methods. Figure 1 displays an example of an empirically derived ERP along with its best and worst DIVA reconstruction. Note that, as was expected, the DIVA classifier learned to emphasize some parts of the input waveforms over others.

Naive Discriminant Learning (NDL)

The NDL model was trained using a slight extension of the NDL algorithm developed by Shaoul, Arppe, Hendrix, Milin, and Baayen (2013). Essentially, this

model can be considered as a two layer network with 550 inputs, corresponding to each sample of the full ERP waveform, and 32 outputs, corresponding to each of the participants who may have generated the waveform. This network was trained using the Danks (2003) equilibrium equations to identify threshold values and weights for above-and below-threshold inputs that should be fed forward to each of the output units, to maximize the activation of the correct output and minimize the activation of the incorrect outputs. The use of these equilibrium equations effectively allows for the weight matrix that would be discovered by iterative discriminative learning across the training examples (e.g., as in back-propagation) to be derived in a single sweep through the corpus. Following training, the threshold values and weights were fixed and were used to generate the predicted outputs for the testing data set. Activation of the output units was then ranked to generate an analogous set of ranking data to that developed for the other machine learning algorithms outlined above.

Results

Identical analysis was performed on the rankings from each classifier. A rank of 1 was considered a highly confident "vote" for that classification, and was given a weight of 1, whereas a rank of 2 was given a weight of .97 and so on, such that a rank of 32 was given a weight of 0. There are two, related, questions of interest when evaluating the accuracy of multi-way classifiers in this manner. First, how often did the classifier make the "correct" classification (rank the correct classification highest)? Second, when the correct classification is not the first ranked classification, how highly is it ranked? This second question quantifies the idea that if, for example, a classifier ranks the correct classification 2nd, that should be considered a more favorable result than if the classifier ranks the correct classification last.

To answer the first question, we asked how often the correct classification was ranked 1 more frequently than any incorrect classification for each subject (e.g., if the correct classification was ranked 2 more often than it was ranked 1 within a particular subject's 100 test exemplars, that subject was considered incorrectly classified); we will refer to this in what follows as the classification accuracy. The classification accuracy for the cross-correlation classifier was 0.90. The classification accuracy for DIVA was 0.89. The classification accuracy for NDL was 0.89. To answer the second question, the mean of the weights assigned to the correct classification for each subject was taken as a measure of the success of the classifier in identifying that subject, regardless of whether ultimately the classifier actually "chose" the correct classification (i.e., ranked it 1 most frequently). In what follows, we will refer to this as the mean rank weight. The mean rank weight for the cross-correlation classifier was .87. The mean rank weight for DIVA was .90. The mean rank weight for the NDL classifier was 0.88. We also calculated the absolute accuracy for each time a trial was well classified (i.e. was correctly ranked 1) for all the 32000 trials. The results were 0.56 for cross-correlation, 0.54 for DIVA and 0.42 for NDL.

The null hypothesis for classification accuracy is that the classifiers are assigning the first rank by chance; meaning that the chance classification accuracy is 1 / 32 = .03. Clearly, all classifiers performed substantially better than chance. To quantify this statistically, we computed the distribution of decision accuracies across 50 000 random permutations of the ranking matrix. We then assigned p-values to the null hypothesis by determining the proportion of random classification accuracies that were higher than the observed classification accuracy for each classifier (a type of approximate randomization test). Similarly, the null hypothesis for rank weight is that all 32 ranks are being assigned by chance. We assigned p-values to the null hypothesis by determining the proportion of mean rank weights in the random 50 000 permutations of the ranking matrix that were higher than the observed mean rank weight. The null hypothesis was rejected for all classifiers, on both measures of accuracy, at $p < .0001$ (the same was true for absolute accuracy).

Discussion

We set out to investigate whether we could accurately identify individuals on the basis of unique features of their neural activity. After advancing the proposition that one cognitive structure likely to be unique to individuals is the detailed organization of semantic memory, we submitted ERP data acquired in a semantic memory task to multiple pattern classifiers: cross-correlation, DIVA, and NDL. All three classifiers were able to classify individual waveforms with a very high degree of accuracy robustly above chance---indeed, performance was near ceiling in most of our analyses, particularly for the training data. The fact that these results are very similar for the three different methods used shows that the data includes robustly identifiable differences across individuals, which can be detected by a variety of methods. It also demonstrates that our cognitive linking premise-- that access to semantic memory is a uniquely individual process-- is at least not entirely defunct as a rationale.

There are numerous avenues of future research advancing our treatment here of the brainprint problem.

As a one example, it would be interesting to analyze the EEG data in single trials to see if the information of whether an acronym is recognized or not--without trial averaging--can be detected by a classifier. Also, correlations between components (e.g., correlations between the N400 and the P2) might provide another source of identifiable variation between individuals.. On the side of signal processing, using a voting scheme between the algorithms, or even between different electrode sites may improve over the accuracy of any single algorithm.. A point to highlight is that the data processed in this work was collected for different purposes. It could be worthwhile to conduct an experiment tailored specifically to generate a different response by a range of users, in order to understand the upper limits of the brainprinting accuracy.

Finally, our success here has implications for the applied use of brainprinting, as for secure and trustworthy authentication of access to sensitive information. There are multiple advantages of brainprinting over traditional biometrics (such as fingerprints and retinal scans). As opposed to traditional methods, brainprinting protects not only the system from unauthorized access, but also the subject from being harmed in order to acquire its biometric feature, as can happen with fingerprints, for example (BBC news: Malaysia car thieves steal finger). Our success here at uniquely identifying individuals even in a dataset not designed specifically for generating maximally unique waveforms indicates that existing computational methods are sufficiently sophisticated to make applied brainprinting feasible, in principle. In future research, we aim to more rigorously explore the theoretical and practical considerations that will allow this work to be of practical use to society.

Acknowledgments

The authors acknowledge members of the Binghamton University Modeling Meeting--especially S. Bhamdeo and T. Raway--for insightful discussion. This research was supported by the Binghamton University Health Sciences Transdiciplinary Area of Excellence (S.L & Z.J), NSF CAREER BCS-1252975 (S.L.), and the European Science Commission MC IIF-627784 (B.C.A.).

References

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, 118(3), 438-481.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford university press.

Chandaka, S., Chatterjee, A., & Munshi, S. (2009). Cross-correlation aided support vector machine classifier for classification of EEG signals. *Expert Systems with Applications*, 36(2), 1329-1336.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of verbal learning and verbal behavior*, 19(4), 450-466.

Danks, D. (2003). Equilibria of the Rescorla-Wagner model. *Journal of Mathematical Psychology*, 47, 109-121.

Jian-feng, H. U. (2010, March). Biometric System based on EEG Signals by feature combination. In Measuring Technology and Mechatronics Automation (ICMTMA), 2010 International Conference on (Vol. 1, pp. 752-755). IEEE.

Khalifian, N. (2013) Life of ERPLE: Developing a silent reading task for use with children and adults. SUNY Binghamton, Binghamton, New York.

Kurtz, K. J. (2007). The divergent autoencoder (DIVA) model of category learning. *Psychonomic Bulletin & Review*, 14(4), 560-576.

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.

Laszlo, S., & Federmeier, K.D. (2007). Better the DVL you know: Acronyms reveal the contribution of familiarity to single word reading, *Psychological Science*, 18, 122-126.

Palaniappan, R., & Mandic, D. P. (2007). Biometrics from brain electrical activity: a machine learning approach. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 29(4), 738-742.

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., & Baayen, H. (2014). The Myth of Cognitive Decline: Non−Linear Dynamics of Lifelong Learning. *Topics in cognitive science,* 6, 5-42.

Raz, N., Lindenberger, U., Rodrigue, K. M., Kennedy, K. M., Head, D., et al. (2005). Regional brain changes in aging healthy adults: general trends, individual differences and modifiers. *Cerebral cortex*, 15(11), 1676-1689.

Shaoul, C., Arppe, A., Hendrix, P., Milin, P., & Baayen, R. H. (2013). ndl: Naive Discriminative Learning. R package version 0.2.14. http://CRAN.R-project.org/package=ndl