# Using N400-Component for Automatic Thought Identification In a Natural Reading Task

Johan Emilsson
Supervised by: Christian Balkenius
.
Lund University Cognitive Science
johan.emilsson.867@student.lu.se

Recent advances in human brain imaging techniques have shown that it is possible to correctly identify thoughts associated with specific categorises of conceptual knowledge. A semantic space is created through word clustering from a corpus and cross-referenced with words read during a natural reading task while at the same time recording EEG. These results show that it is possible to correctly classify a words category 25% of the time (out of 12 possible categories) based on neural activity within the N400 component with the help of a artifical neural network. Results also visualises the ANN's black-box behaviour in which electrodes it deem most contributing when classifying. These results contribute to the discussion on how the semantic memory is stored in the brain; modularly or connectionisticly, and also how far it is possible to subcategories thoughts and still correctly identify them with brain imagining techniques.

**Keywords:** *EEG; N400; ANN; Artificial Neural Network; Word cluster; PCA; Semantic space; categorisation; reading task*

Humans have a fundamentally important cognitive ability to create and coherently understand meaning. Each experience is unique and to keep up with the high cognitive load this creates, an effective mental storehouse for doing sorting and association is needed of what we "know" about actions, events, people, places, and things, including words, stored in the brain (Kutas & Federmeier, 2011). Understanding how humans perceive the world has just until recently been a bottom-up process. With increased computational power and more elaborate data-driven models to analyse data from neuroimaging technique (fMRI, EEG) it's been possible to identify independent thoughts or thoughts connected to a larger category or concept. This technique has been used to identify important structures in thoughts coding for *nouns* (Mitchell et al., 2008; Just et al., 2010), *semantic category* (Murphy & Poesio, 2010), *concepts* (Gerven & Simanova, 2010), and *what drawing was viewed* (Shinkareva et al., 2008).

## Background

### *Semantic memory*

Semanic memory, memory of meaning, is a theorized neural structure where semantic knowledge is organized and stored (Canessa et al., 2008). There is a constant flow of information from the surrounding area to be processed by the cortex and decide how to act upon that processed information. This is called the Perception-Action Cycle; information from the sensory organs integrates with ones actions and both are means to each other, creating an everlasting cycle of informational exchange between those two (Caramazza & Mahon, 2003). This cycle, for just mentioned reason, has thus lead to specific neural structures being created due to an evolutionary need to be able to process perceptual and conceptual categories of objects. Opposing side suggest rather that semantic memory is part of not a specialised module distinct from the rest of cognition, but more part of broad human cognitive skills (Dunn et al., 2011). It could, from this point of view, be that what we call *semantics* is just the way that a large collection of brain cells works, the evolvement of a functional general purpose cognitive 'cloud' - an aspect of thought itself.

*Neural source of the N400-component*. With recent increase in usage of N400 as an index of semantic memory has it also become of interest to localize the source of this activity. This is to help understand, not just what brain areas that are involved, but also when and how they might contribute (Van Petten & Luka, 2006). Results show a widely distributed set of active brain regions where N400-component been measured. These include a source in the anterior medial temporal lobe, in middle and superior temporal areas, inferior temporal areas and, to some less consistency, prefrontal areas (Lau et al., 2008). Just mentioned sources of the N400-component have been localized in both hemispheres, but the results have been shown to be a bit stronger in the

A sincere thank you to Petter Kallioinen, Christian Balkenius, Thomas Strandberg, Frida Svärdby, and Rasmus Bååth for contributing to my thesis with thoughts, ideas, suggestions, corrections and help with coding.

left hemisphere (Vigneau et al., 2011). This goes hand-in-hand with recent advances in connectionist versus modularity research that a distributed network presumed integral to semantic memory storage and processes exists (Burianova et al., 2010) and not as previously thought localised to a single static source (Fuster, 2000; Calabretta & Parisi, 2005). So, the contemporary consensus in the field of neuroimaging is that data point to a multimodal semantic system (Price, 2010). Scalp reading on the N400 suggests that it's a wave of activity starting roughly 250 ms from stimuli onset in the posterior half of the superior temporal gyrus, propagating itself forward and ventrally to the left temporal lobe (365 ms) and thereafter to the right anterior temporal lobe and to both frontal lobes (370-500 ms). This opens up the question of the functions of the dynamic neural system and what the N400 reflects (Binder et al., 2009).

*Semantic representation*. Semantic representations are statisical models emulating semantic knowledge and is possible to create in an artificial way by statistical means. These models are often based on a predefined context to operate inside and measure words fitness in a context by number of times it occurs in that given context. This is done by creating a large matrix with word occurrences in every possible context. Context in this case equals to a document, article or text inside a larger corpus. This could lead to words that typically occurs in the same context is not seen as something, by the model, that fit together. Examples on such models are *Latent Semantic Analysis* (Landauer & Dumais, 1997), and *Hyperspace to Analogue Language* (Lund & Burgess, 1996).

### Pattern recognition with data-driven models

There is a high number of different data-driven classification algorithms that has been applied to different fields of research (Lotte et al., 2007), for example with the help of *Artifical Neural Network* (ANN), *Support Vector Machine* (SVM), and *Self-Organizing Maps* (SOM). What they have all got in common is that they are data-driven; they create models based on incoming data. This creates a problem and that is the emergence of a so called "black-box" behaviour, problem to follow the way of thought for input that the model deems more important than other sources of information. Solutions to circumvent this "black-box" behaviour has been proposed in previous research (Mika et al., 1999; Muller et al., 2001). The problem from a human perspective is that the data-driven models find high dimensional solutions and those are impossible for a human being to interpret and understand correctly. The solution is to reduce number of dimensions. One suggestion in how to do this is by using a principal component analysis (PCA) or one of the many varieties of it (Jimenez et al., 2009).

### The focus of this thesis

This is an explorative study investigating the distribution of the semantic memory, with the help of data-driven models. The assumption that is made is it's a necessity to use data-driven models to classify and find correlations in semantic representations due to its high dimensionality solutions. The question is too complex for a strict bottom-up way of thought and that is why a classifier is needed to understand the connections, so to reduce complex connections to classifiable connections. The question to be answered is how the semantic memory is distributed and also contribute in the debate whether the semantic memory is modular or connectionistic. Also of interest is if there is a correlation between semantic clusters in a written text and neural activation. Therefore, it is hypothesis that word clusters or categories with high internal similarity also have a higher neural distinctiveness resulting in a high percentage of correct classifications.

*Limitations*. Limiting factors would mainly be to what extent generalizability is possible, since the experimental design is based mainly on one test subject (author). Other limiting factor, for good or bad, is that only one component is to be used during classification. Using only one component focuses on only one aspect of how one interprets a word, while using more components would hinder ease of analysis due to increased number of factors involved.

## Method

### Subject

One participant (author, J.E, age 26) was used during the data acquisition. The participant is at good health, adequate eyesight - after correction, left-handed (determined by tooth brushing hand), and has Swedish as native language.

### Stimuli

*Corpus*. The size of the complete corpus that was used in this study amounted to a total of 740 000 words. It was constructed from Swedish newspaper articles that was published during 2005 and included various themes; sport-, world-, local- and regional news, that totalled at 28 000 different articles. The stimulus material was sampled from this corpus trough a randomisation procedure. Complete articles was collected from the corpus through this procedure until around 20 000 words was sampled. A segment of text that was under 40 words or more than 400 words was filtered away. Upper and lower limit for the filter was decided arbitrarily. Filtering was made because articles less than 40 words wouldn't give enough semantic contexts for later analysis and articles containing more than 400 words seized too much space, so there would not have been the same diverse distribution among the article themes. Thus by this reasoning the generated sample are reading from the whole corpus when it comes to the same context. The size of the stimulus material is motivated for the reason that it holds high ecological validity, by simulating natural text reading. The text used in the experiment is not corrected for word length or word frequency bias.

### Procedure

Experimental construction was completed with E-prime v.1.2 (Schneider, Eschman, & Zuccolotto, 2002) and modulated to synchronise with the EEG recording equipment, EGI

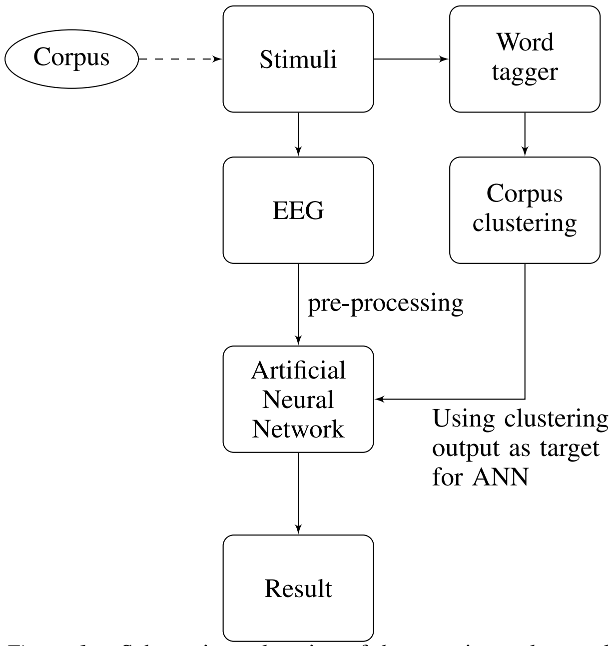


*Figure 1*. Schematic explanation of the experimental procedure, analysis and theorised results. Corpus gets pre-processed to sample the stimuli for the next step, experiment. Stimuli are presented for the participant while at the same time EEG is recorded. Presented stimuli are later categorized automatically with SenseCluster. The output from the categorization is after that cross-referenced with words having EEG-data collected on them. Those clusters with enough EEG-data is after that trained in the artificial neural network, ANN. The theorised output will correspond to important areas that helps the ANN to classify what cluster a word belongs to and also evaluate how well the automatic categorization corresponded to collected EEG-data.

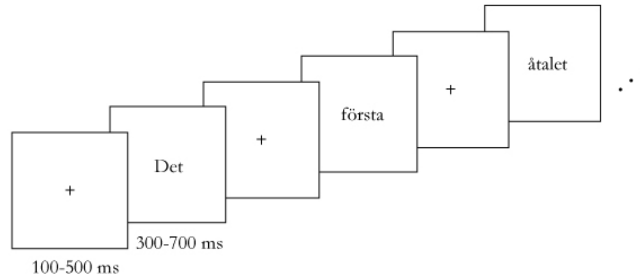


*Figure 2*. This is the experimental setup that was used the experiment, where a participant was presented with 20000 Swedish words. Each fixation and stimuli period were randomized in the interval of 100-500 ms (fixation) to 300-700 ms (stimuli). Fixation came prior to the stimuli event. White text for fixation point and stimuli were presented on a black background, but is in the figure inverted for ease of reading.

Netstation. The EEG net that was used was of model HydroCel Geodesic Sensor Net and had 128 electrodes plus reference electrode. The participant was placed in a desk chair in front of a computer screen, with the instructions to read the presented text just like when you read in a book, but with the inclusion to avoid, if possible, to blink. The EEG equipment was mounted shortly after the instructions had come to an end.

In front of the participant words was presented, one at a time, from the randomized collected articles in the text corpus, on a computer screen, while at the same time being connected to the EEG equipment. Words are presented in a standardized manner; fixation cross and then stimuli (see figure 2). This approach, to collect single-trial EEG for 20000 words, is divided into 14 sessions (roughly 30 minutes for each session). At each new session recalibration of the equipment is done, to see that the impedance for the electrodes are below an acceptable threshold and to give the participant a time to regain some vigilance, since there is no way to pause a session. The time from onset of a fixation cross until stimuli shows up is between 100-500 milliseconds and for stimuli 300 to 700 milliseconds. After that the cycle is repeated. The reason for the randomized time jitter is to avoid an expectation effect from participant. See figure 1 for an overview of the experiment design.

## Data analysis

*Word clustering*. Word clustering (e.g.: synonym finding) is done by SenseClusters (Kulkarni & Pedersen, 2005; Pedersen et al., 2005, 2006; Pedersen, 2008), native methodology. Word or *bigram* clustering is the result from a second order representation. Second order representation creates a vector for each context that indicates which words occur with the words in that context. By creating a word by word matrix from bigrams, a context in which those words co-exist is created. This feature matrix is reduced by a *Singular Value Decomposition*, SVD (Golub & Kahan, 1965) prior to clustering. A reduction factor, K, set at 300 is used, consequently reducing the matrix dimension of the feature space to K. The scaling factor used for the SVD procedure is set at 10. This implicates the scaling factor for which reducing feature space dimensions such that feature with N dimensions is reduced down to N/K. This reduction replaces the word by word matrix with a corresponding vector, so in the end, are all of these vector averaged together to represent the context. This averaged vector is the centroid of all the word vectors that make up the context. Words are by this method clustered together with words with which they co-occur. This differ in methodology from Latent Semantic Analysis, in which clusters features is based on the context in which they occur. As input a senseval-2 formatted file was created out of the corpus. The 75 most frequent words in the corpus was used as stop words, since most of those words do not contain useful information regarding the clustering. Clustering method that was used was a *repeated bisections with a K-way refinement* and words that co-occurred less than 5 times together in a window of 2 words got filtered away. Criteria function for clustering was an CLUTO $I_2$ function (Karypis, 2002; Zhao & Karypis, 2004) using cosine as a similarity measurement. The $I_2$ function represents each cluster by its centroid vector and searches for the solution that maximizes the similarity between document and the centroid of the cluster that it is assigned to. A pre-decided numbers of clusters was used, 100, and the reason for that was a shortage of computational power for cluster size optimization.

*Event-Related Potential*. During pre-processing a Finite Impulse Response (FIR)-filter that encompass 0.3 to 40 Hz, passband and stopband gain adjusted to 99.0 respectively 1.0 %, are applied to the datasets. Artefact detection marks segment bad if it contains more than 10 bad channel, eye blinks or eye movement. Datasets are baseline corrected. Baseline begins at 0 milliseconds in each segment and is 100 milliseconds long. Words with EEG recorded to it are grouped into its corresponding cluster. A minimum of 150 words with EEG is decided to be the cut-off level for inclusion to the next step in the analysis. The cut-off number is based on the assumption that it is necessary to have at least 150 trials per variable, in this case clusters, to be able to make a good ERP average with a high signal to noise ratio (Luck, 2005). Words that do not belong to any cluster meeting the criteria of 150 trials are filtered away. A butterfly plot is used to identify the N400-components time interval (380-450 ms). This means that for each time interval 18 measurements is taken (once every fourth millisecond).

*Artificial neural network*. Input into the Artificial Neural Network, ANN, is the result from the Word clustering and ERP filtering and works as a supervising template for the classification procedure. For this pattern recognitions task a two-layer feed forward back propagating network with sigmoid output neurons, ANN, is used. Implemented using Mathworks Neural Network Toolbox (Demuth & Beale, 1998). This corresponds to a supervised artificial neural network, since it is guided by a desired outcome, target class. Input is trained to classify according to target classes (word clusters). Input nodes, in this case, equals to each of the 128 EEG electrodes. Number of neurons in the hidden layer is something that is adjustable and depends on selected task. In this case 10 neurons in the hidden layer were found optimal. Data division are divided into training validation and testing set, with the following percentage each: 70%, 15% and 15%. Testing set is not used during training to avoid biasing due to previous exposure. Testing ends when it has reached a predetermined gradient, in this case $10^{-6}$ or six consecutively correct classifications from the validation set. Performance for each round is determined by mean squared error, which is the average squared difference between outputs and targets. Lower is better and zero means that no errors have been made.

Four word clusters are selected based on a randomised picking for further analysis to prove the concept of using CATPCA (see next section) as a tool to interpret and visualise an Artificial Neural Networks black-box behaviour. Those four word clusters are run against each other two at a time until every possible combination is done. To do every combination a total of 6 different cluster combinations is needed; 1-2, 1-3, 1-4, 2-3, 3-4, and 3-4. Each of the 6 combinations plus one with every word cluster is reiterated 20 times to be able to get a reliable average from the ANN's weights necessary for the next step.

*Categorical principal components analysis (CATPCA)*. Categorical principal components analysis (CATPCA) is used on the resulting ANN's weighs created after each run. This is to identify the underlying components of EEG-data that the ANN's based its weights on, while maximizing the amount of variance accounted for in relevant electrodes. The benefit with CATPCA is that it does not assume a linear relationship, which suits analysis of the n-dimensional weights well. Previous studies have shown that it is possible to use CATPCA for dimensionality reduction during data mining well (Jimenez et al., 2009). Each configuration (6+1) weights are discretised into 10 ordinal categories. This is hypothesised to correspond with a Gaussian distribution of the 10 neurons used earlier in the ANN. Each electrodes explained variance, according to the CATPCA in two factor latent model, is inserted into EEGlab (Delorme & Makeig, 2004), to visualise the relationship between each electrode on the scalp. To see if there exists any variation between what the ANN deem importation during categorisation each of the four word clusters EEG is combined with the three times each word cluster is run with one of the other clusters (for word cluster two: 1 vs 2, 2 vs 3, and 2 vs 4). EEGlab is used to account for each electrode XYZ-location in a Cartesian coordinate system.

# Results

## Word clustering

From the word clustering, 12 clusters matched the criteria of having at least 150 words with EEG out of the 100 different clusters. This is summarised in table 1. Results from the 100-way word clustering, done by SenseCluster, has a average similarity value, $I_2$, of 4,09e+03 in a total of 8469 different words. It is highly likely to agree that most of the content correctly belong to it assigned cluster by visual inspection, but there are discrepancies that is not likely to occur if the categorisation had been done by a human. These categorizations errors occur due to words that are written together show a statistical relationship, but could in reality be far from the truth. It is also possible that this error also creates an increased noise during the ANN categorisation, resulting in a lowered overall score. The theme of the clusters is subjectively decided by the author after reviewing each clusters content. Most clusters are based on word class, but far from all. It is possible to identify several semantically related and correct categories created by SenseCluster; for example: *titles people hold that are involved with politics in a Swedish municipality*, *school*, *Muslim world*, *industry* and *economy* but not included in the final analysis due to the 150 word criteria.

Similarity measures between and within categories is shown in table 2. They show the degree of similarity between and within categories with help from the cosine function. Most of the clusters have a high internal similarity when

Table 1
*Summarising table from the word clustering applied to the corpus and then co-referenced with the collected words during the EEG experiment.*

| # | N in corpus | N with EEG | Interpreted Theme | Descriptive Words[a] |
|---|---|---|---|---|
| 1 | 99 | 385 | Adverb | allt, göra, mig, aldrig, bli, alltid, säga, ännu, blev, betydligt |
| 2 | 131 | 329 | Noun | vilket, polisen, personer, familjen, året, HV, person, Landskrona, slut, kvinnan |
| 3 | 119 | 294 | Politics & democracy | du, detta, först, därför, Sverige, samtidigt, själv, domstolen, ni, regeringen |
| 4 | 104 | 292 | Verb (communication) | Sydsvenskan, fick, måste, kunna, skall, började, låta, börjar, vän-tas, hoppas |
| 5 | 117 | 264 | Law & criminality | allvarligt, skäl, inträffade, skadades, brand, midnatt, skäligen, släcka, häktad, anhållen |
| 6 | 95 | 263 | Abstract noun | själva, oss, dag, honom, tiden, fel, dem, igen, matchen, ingenting |
| 7 | 1883 | 216 | Name | Zlatan, Ilmar, Universitetssjukhuset, Los, statsministrar, power, Sven, Turning, Barack, Telia |
| 8 | 107 | 190 | Verb, present tense active | genom, gjorde, Karolinska, ur, lämna, fly, tillbaka, ökat, minska, organiserade |
| 9 | 81 | 181 | Location & time | senaste, halv, minst, mellan, kl, ytterligare, ungefär, cirka, klockan, nästan |
| 10 | 106 | 180 | Verb, preterite | ibland, se, ät, ge, gav, kalla, sköt, hjälpa, bakom, gett |
| 11 | 109 | 165 | Small numbers (1, 2) | 5, 4, 3, 6, 2, 8, 30, 7, 20, 40 |
| 12 | 64 | 150 | Number names (one, two) | flera, några, tre, tio, fyra, fem, par, sex, upprepade, sju |

[a]First ten words in each cluster

placed in context of the semantic space and all have a low external similarity. Those categories that have a lower internal similarity (foremost: *adverb* and *name*) are also word classes that are hard to define. Adverb in the Swedish language is a word class category that is used when a word is hard to place in any other of the word classes, and is considered to be something of a ”catch-all“ category. Names are also a word class that is hard to categorise without a narrowed context.

## *Classification with artificial neural network (ANN)*

With a 12 category solution it is possible by the artificial neural network to classify correctly, on average, 25,4±2,01% of the time, compared to 8,3% if it had been by chance. Thus saying it is possible to identify thoughts from EEG up to a certain degree. A average run with a 12 category solution can be seen in table 3. Categories 1 through 7, with the exception of category 3, stands out from the remaining categories because they all have a higher percentage of correct classifications then the average score and small confidence intervals (see figure 3 for 95% CI). One possible reason why the top performing categories are just top performing in the ANN could stem from the fact that they mostly consist of nouns and verbs, and more or less adverb and names since they are subclasses to nouns and verbs. Nouns and verbs are also the most frequent word classes in the Swedish language. Previous research have shown that these two word classes are represented in distinct neural networks and is highly important in how humans perceive and categorises the world, compared to the other more subtle categories (Perani et al., 1999; Matzig et al., 2009). Reasons for the results not being

Table 2
*Internal and external clustering quality statistics. $Sim_{int}$ display the average internal similarity between the objects of each cluster and $Sim_{ext}$ display the average similarity of the objects of each cluster and the rest of the objects. A high $Sim_{int}$ and low $Sim_{ext}$ is the most desirable result*

| # | Interpreted theme | $Sim_{int}$ | $SD_{int}$ | $Sim_{ext}$ | $SD_{ext}$ |
|---|---|---|---|---|---|
| 1 | Adverb | 0,24 | 0,11 | 0,00 | 0,00 |
| 2 | Noun | 0,52 | 0,17 | 0,01 | 0,01 |
| 3 | Politics & democracy | 0,37 | 0,12 | 0,01 | 0,01 |
| 4 | Verb (communication) | 0,27 | 0,10 | 0,00 | 0,00 |
| 5 | Law & criminality | 0,35 | 0,15 | 0,01 | 0,01 |
| 6 | Abstract noun | 0,53 | 0,17 | 0,01 | 0,01 |
| 7 | Name | 0,00 | 0,00 | 0,00 | 0,00 |
| 8 | Verb, present tense active | 0,46 | 0,16 | 0,00 | 0,00 |
| 9 | Location & time | 0,22 | 0,10 | 0,00 | 0,00 |
| 10 | Verb, preterite | 0,56 | 0,16 | 0,01 | 0,01 |
| 11 | Small numbers (1, 2) | 0,62 | 0,12 | 0,00 | 0,00 |
| 12 | Number names (one, two) | 0,30 | 0,13 | 0,00 | 0,00 |

any higher than 25% could be the result from a high ratio of noise in the EEG data. Another possible villain could be that the word clustering didn't make sensible clusters, consequently creating a cluster that represents human categorisation. A third possibility would be that the N400-component

Table 3
*Confusion matrix from an average run with a 12 cluster solution in the artificial neural network. Numbers in the matrix represents samples from the N400 component in ANN run number 5, as this equals an average run. One sample equals a measurement taken once every fourth millisecond. Grey cells with bold text represent correct classifications. N=20.*

| **#** | **Output Cluster** | **Distribution matrix** | | | | | | | | | | | | **Correct$_{avg}$** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | Adverb | **372** | 126 | 135 | 134 | 143 | 155 | 124 | 102 | 102 | 96 | 102 | 107 | 25,1% |
| **2** | Noun | 54 | **289** | 90 | 80 | 74 | 75 | 90 | 66 | 80 | 52 | 55 | 59 | 26,6% |
| **3** | Politics & democracy | 64 | 103 | **248** | 71 | 97 | 64 | 74 | 81 | 61 | 63 | 53 | 34 | 21,6% |
| **4** | Verb (communication) | 72 | 61 | 74 | **225** | 47 | 50 | 48 | 36 | 63 | 51 | 48 | 49 | 29,0% |
| **5** | Law & criminality | 86 | 63 | 45 | 50 | **172** | 39 | 35 | 52 | 46 | 42 | 34 | 37 | 26,2% |
| **6** | Abstract noun | 62 | 37 | 46 | 49 | 39 | **180** | 40 | 37 | 43 | 38 | 28 | 29 | 27,4% |
| **7** | Name | 13 | 20 | 29 | 15 | 16 | 20 | **31** | 15 | 19 | 12 | 13 | 20 | 27,7% |
| **8** | Verb, present tense active | 18 | 16 | 21 | 17 | 39 | 6 | 21 | **68** | 9 | 11 | 16 | 6 | 21,2% |
| **9** | Location & time | 2 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | **2** | 0 | 0 | 0 | 13,8% |
| **10** | Verb, preterite | 28 | 4 | 30 | 26 | 12 | 11 | 20 | 11 | 15 | **85** | 13 | 21 | 22,5% |
| **11** | Number names (1, 2) | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | **3** | 0 | 22,5% |
| **12** | Number names (one, two) | 7 | 18 | 11 | 4 | 8 | 10 | 13 | 7 | 7 | 5 | 9 | **56** | 22,8% |
| | **Correct classfications** | 47,8% | 29,2% | 24,0% | 23,5% | 26,6% | 29,5% | 9,2% | 4,3% | 0,4% | 28,6% | 0,8% | 23,4% | **25,3%** |
| | | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **Total** |
| | | **Target Cluster** | | | | | | | | | | | | |

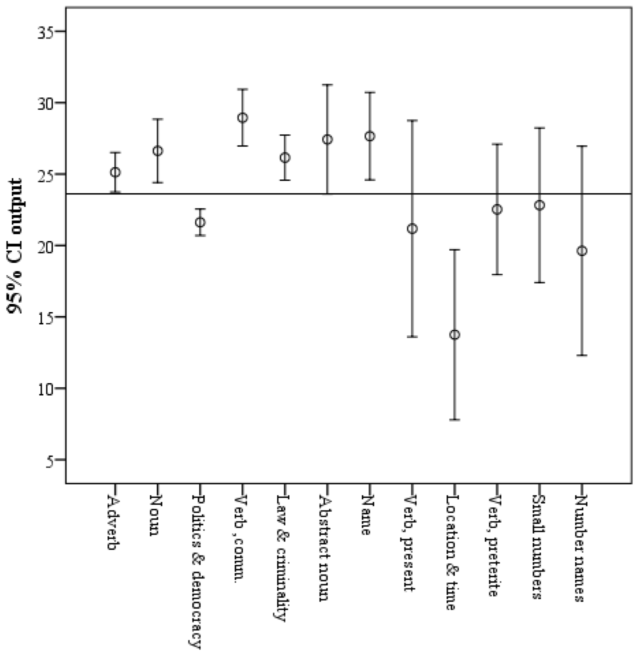


*Figure 3.* A figure with 95% Confidence Interval for each of the 12 categories showing performance during classification in the ANN. Dividing line, at 25,4%, equals total mean. Categories confidence interval is increasing as a function of number of words with EEG connected too, therefore a high number of words equals a small confidence interval.

only contain so much information about each word and its category as it have shown in this study, and to further increase a positive outcome more components would be needed; for example P300, ELAN and N200 components.

There exist no correlation between internal similarity from the word clustering and averaged results from the ANN categorisation. This is contrary to the proposed hypothesis that word cluster with a high internal similarity would also show a high percentage of correct classification due to a distinct neural pattern.

By looking at the table 3 it is possible to identify categories more prone then others to get misclassified into another category. Reason for this might be that there is an overlap in semantic representation in the distribution of semantic memory between two categories or might just be a random anomaly, due to small sample size and a high noise ratio. Two categories have been identified to misclassify into another category, as they stand out from the rest of the categories in number of misclassified samples attributed to them. Samples from the cluster *law and criminality* do often get misclassified by the ANN as a *adverb* and also *verb, preterite* with *law and criminality*. *Location & time* is an example that did poorly on both clustering and during the classification run. Seeing what other words this cluster contains than just location and time related word there are, one notice a high number of out of this specific context words; mostly numbers, but also words related to everyday life.

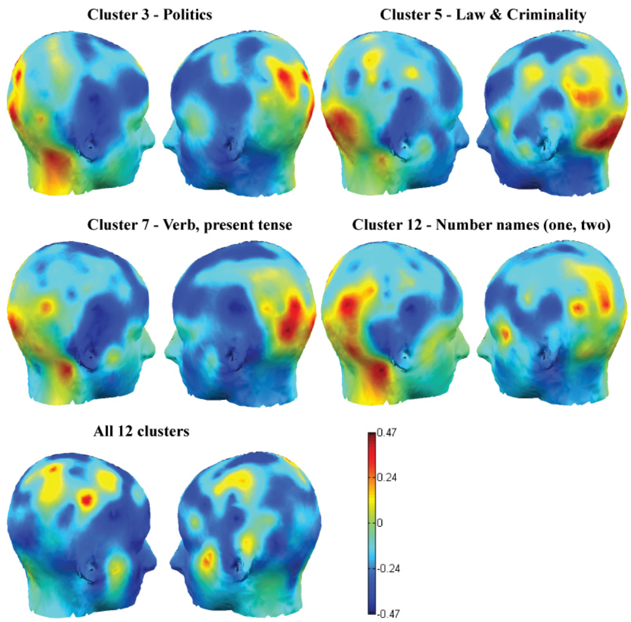


*Figure 4*. Output from the CATPCA, visualised trough EEGlabs' scalp 3D-plotting function, over which electrodes code for most explained variance in the N400 component during classification with the ANN. Figure is divided into four specific categories and one (bottom left) is the output from all 12 categories at the same time. Red colour indicates most important electrodes for the ANN to base its classification on.

*Visualising black-box behaviour*. Since the artificial neural network is data-driven method, it is hard to follow its way of though, creating so called black-box behaviour without insight in its inner workings. Combining results from CATPCA and EEGlabs' plotting functionality, this incognito behaviour is partially overcome (see figure 4). Output in figure 4, show that electrodes the ANN deem important when classifying is distributed posterior, in the vicinity of the occipital and parietal lobe. These results may be due to the fact that the experiment was a visual task and important areas for ANN are also areas that is involved with perception. The spotted pattern that emerged from the CATPCA might show that ANN uses all available channels, but neighbouring channels is very similar thus making them redundant.

## Conclusion

The results presented in this thesis adds valid proof to what previous research also suggests about thought identification through data-driven models and that is it is possible. It has been shown that EEG contains enough discriminate information so that an ANN can discriminate a words belonging between 12 categories. It has also been shown that it's possible to visualise the inner working in a data-driven model that usually only exhibit a black-box behaviour. Since ANN is not deterministic it is not possible to say these results are universal. Further runs with the ANN with different configurations will be needed to determine that. This also has to do with the sampling and initial starting position that changes for each round run. Hopefully it's been avoided with running each interval 20 times.

### *Future research*

It would be possible to improve and tighten this experiments design further. By limiting number of word classes, except for nouns, it would then be able to create a semantic space comprised of ”relevant“ semantic structures and hopefully correlate better with structures in the semantic memory. Another version of this experiment, to further confirm its claim and field of research, would be to have a auditory stimuli instead. This would be to see if the location of the increased areas of interest for the ANN and deduced into understandable results by the CATPCA would change from perception related areas to auditory areas, thus increasing the validity of applied method.

### *Implications*

The implications of this type of research, would in the far end be a life saver for those that lost the possibility to communicate with the outer world and ease the implementation of tools to people with different handicap; pedagogical tools; and in the close vicinity, improving the understanding in how categorisation works and how categories is created and stored in the semantic memory. A note of caution in drawing further conclusions than the just above mentioned is that a language does not just contain twelve categories, as used in this thesis. Language is highly complex, so we can't expect anyone in the near future to go prodding around inside a brain reading every thought word or stray thought.

## References


Binder, J., Desai, R., Graves, W., & Conant, L. (2009). Where is the semantic system. *A critical review and meta-analysis of*, *120*, 2767–2796.

Burianova, H., McIntosh, A., & Grady, C. (2010). A common functional brain network for autobiographical, episodic, and semantic memory retrieval. *Neuroimage*, *49*(1), 865–874.

Calabretta, R., & Parisi, D. (2005). Evolutionary connectionism and mind/brain modularity. *Modularity. Understanding the development and evolution of complex natural systems*, 309–330.

Canessa, N., Borgo, F., Cappa, S., Perani, D., Falini, A., Buccino, G., et al. (2008). The different neural correlates of action and functional knowledge in semantic memory: An fmri study. *Cerebral Cortex*, *18*(4), 740.

Caramazza, A., & Mahon, B. (2003). The organization of conceptual knowledge: the evidence from category-specific semantic deficits. *Trends in Cognitive Sciences*, *7*(8), 354–361.

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.

Demuth, H., & Beale, M. (1998). Neural network toolbox users guide. *The MathWorks Inc*.

Dunn, M., Greenhill, S., Levinson, S., & Gray, R. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, *473*(7345), 79–82.

Fuster, J. (2000). The module: Crisis of a paradigm. *Neuron*, *26*(1), 51–53.

Gerven, M. van, & Simanova, I. (2010). Concept classification with bayesian multi-task learning. In *Proceedings of the naacl hlt 2010 first workshop on computational neurolinguistics* (pp. 10–17).

Golub, G., & Kahan, W. (1965). Calculating the singular values and pseudo-inverse of a matrix. *Journal of the Society for Industrial and Applied Mathematics: Series B, Numerical Analysis*, *2*(2), 205–224.

Jimenez, R., Gervilla, E., Sese, A., Montano, J., Cajal, B., & Palmer, A. (2009). Dimensionality reduction in data mining using artificial neural networks. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *5*(1), 26.

Just, M., Cherkassky, V., Aryal, S., & Mitchell, T. (2010). A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PloS one*, *5*(1), e8622.

Karypis, G. (2002). *Cluto-a clustering toolkit*. Minnesota Univ Minneapolis Dept of Computer Science.

Kulkarni, A., & Pedersen, T. (2005). Senseclusters: unsupervised clustering and labeling of similar contexts. In *Proceedings of the acl 2005 on interactive poster and demonstration sessions* (pp. 105–108).

Kutas, M., & Federmeier, K. (2011). Thirty years and counting: Finding meaning in the n400 component of the event-related brain potential (erp). *Annual review of psychology*, *62*(1).

Landauer, T., & Dumais, S. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211.

Lau, E., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics:(de) constructing the n400. *Nature Reviews Neuroscience*, *9*(12), 920–933.

Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., & Arnaldi, B. (2007). A review of classification algorithms for EEG-based brain–computer interfaces. *Journal of neural engineering*, *4*, R1.

Luck, S. (2005). *An introduction to the event-related potential technique*. The MIT Press.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods*, *28*(2), 203–208.

Matzig, S., Druks, J., Masterson, J., & Vigliocco, G. (2009). Noun and verb differences in picture naming: Past studies and new evidence. *Cortex*, *45*(6), 738–758.

Mika, S., Scholkopf, B., Smola, A., Muller, K., Scholz, M., & Ratsch, G. (1999). Kernel PCA and de-noising in feature spaces. *Advances in neural information processing systems*, *11*(1), 536–542.

Mitchell, T., Shinkareva, S., Carlson, A., Chang, K., Malave, V., Mason, R., et al. (2008). Predicting human brain activity associated with the meanings of nouns. *science*, *320*(5880), 1191.

Muller, K., Mika, S., Ratsch, G., Tsuda, K., & Scholkopf, B. (2001). An introduction to kernel-based learning algorithms. *Neural Networks, IEEE Transactions on*, *12*(2), 181–201.

Murphy, B., & Poesio, M. (2010). Detecting semantic category in simultaneous eeg/meg recordings. In *Proceedings of the naacl hlt 2010 first workshop on computational neurolinguistics* (pp. 36–44).

Pedersen, T. (2008). Computational approaches to measuring the similarity of short contexts: A review of applications and methods. *Arxiv preprint arXiv:0806.3787*.

Pedersen, T., Kulkarni, A., Angheluta, R., Kozareva, Z., & Solorio, T. (2006). An unsupervised language independent method of name discrimination using second order co-occurrence features. *Computational Linguistics and Intelligent Text Processing*, 208–222.

Pedersen, T., Purandare, A., & Kulkarni, A. (2005). Name discrimination by clustering similar contexts. *Computational Linguistics and Intelligent Text Processing*, 226–237.

Perani, D., Cappa, S., Schnur, T., Tettamanti, M., Collina, S., Rosa, M., et al. (1999). The neural correlates of verb and noun processing. *Brain*, *122*(12), 2337.

Price, C. (2010). The anatomy of language: a review of 100 fmri studies published in 2009. *Annals of the New York Academy of Sciences*, *1191*(1), 62–88.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-prime: User's guide*. Psychology Software Inc.

Shinkareva, S., Mason, R., Malave, V., Wang, W., Mitchell, T., & Just, M. (2008). Using fmri brain activation to identify cognitive states associated with perception of tools and dwellings. *PLoS One*, *3*(1), e1394.

Van Petten, C., & Luka, B. (2006). Neural localization of semantic context effects in electromagnetic and hemodynamic studies. *Brain and Language*, *97*(3), 279–293.

Vigneau, M., Beaucousin, V., Herve, P.-Y., Jobard, G., Laurent, P., Crivello, F., et al. (2011). What is right-hemisphere contribution to phonological, lexico-semantic, and sentence processing? insights from a meta-analysis. *Neuroimage*, *54*(1), 577–593.

Zhao, Y., & Karypis, G. (2004). Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning*, *55*(3), 311–331.