Classification from Generation: Recognizing Deep Grammatical Information During Reading from Rapid Event-Related fMRI

Tali Bitan[1,2], Alex Frid[3], Hananel Hazan[4], Larry M. Manevitz[3], Haim Shalelashvili[3], Yael Weiss[5]

[1]Psychology Department,University of Haifa, Israel
[2]Department of Speech Pathology, University of Toronto, Canada
tbitan@research.haifa.ac.il
[3]Neurocomputation Laboratory and the Department of Computer Science, University of Haifa, Israel
{alex.frid, haimshalev}@gmail.com, manevitz@cs.haifa.ac.il
[4]Network Biology Research Laboratory, Technion, Haifa, Israel
hananel@hazan.org.il
[5]Brain Development Laboratory, University of Texas, Austin
ylweiss@gmail.com

***Abstract*—A novel fMRI classification method designed for rapid event related fMRI experiments is described and applied to the classification of loud reading of isolated words in Hebrew. Three comparisons of different grammatical complexity were performed: (i) words versus asterisks (ii) "with diacritics versus without diacritics" and (iii) "with root versus no root". We discuss the most difficult task and, for comparison, the easiest one. Earlier work using more standard classification techniques (machine learning and statistical) succeeded fully only in the simplest of these tasks (i), but produced only partial results on (ii) and failed completely, even on the training set on the deepest task (iii).**

The method performs a "best match" between pre-processed data and computing a full library of artificially generated examples. The method involves a deconvolution of the rapid events on the data and performing a convolution on the generated data. The best-match is performed over all "words" constructed by convolving the response functions of each value of each event performed in a "windowed" sequence. This is accomplished separately for all voxels and then a voting procedure defines the outcome.

Using the same feature selection (ANOVA) as in the earlier methods, (i) there is a dramatic increase in the accuracy rate for the third (most difficult task) on the intra-run level (88%) as well as the first task (ii) Unlike the earlier methods training and testing over all runs (within subject) achieves a significant level of classification (64% accuracy) for the training set. This shows the information for this "deeper" cognitive task can in fact be extracted from the fMRI information.

***Index Terms*—Functional magnetic resonance imaging (fMRI), Multivoxel pattern analysis (MVPA), Machine Learning, Pattern Matching, Neural Networks, Cognitive Processing, Classification**

I. Introduction

Functional Magnetic Resonance Imaging (fMRI) is a technique for determining which parts of the brain are activated by different types of physical sensation or cognitive activity, such as sound, the movement of a subject's arm, emotion etc. fMRI provides indirect measurement of the neural activity through the change in Blood Oxygen level, BOLD (Blood Oxygenated Level Dependent Contrast).

Use of multivoxel pattern analysis (MVPA) to predict the cognitive state of a subject during task performance has become a popular focus of fMRI studies. The input to these analyses consists of activation patterns corresponding to different tasks or stimulus types [1]. These activation patterns are fairly straightforward to calculate for block design trials or slow event-related designs, but for rapid event-related designs the evoked BOLD signal for adjacent trials will overlap in time, complicating the identification of the signal unique to specific trials. Rapid event-related designs are often preferred because they allow for more stimuli to be presented and subjects tend to be more focused on the task.

Three comparisons were performed in a previous work [3] for stimuli presented in reading aloud task: (1) Hebrew words compared to string of asterisks; (2) Hebrew words with diacritic marks and words without diacritic marks; (3) Hebrew words with roots and templates and Hebrew words without roots. (Words with roots can be segmented into a root and template, while words with no roots cannot.) In this work we were primarily interested in the performance on task 3.

Tali Bitan and Yael Weiss tried to use SPM, Statistical Parametric Mapping, tool set (SPM, http://www.fil.ion.ucl.ac.uk/spm/, [2]) to achieve significant separation between the activation areas of each condition. Significant separation was achieved for the first two comparisons, with only weak results for the whole brain analysis of the third, and most interesting, comparison of words with and without a root.

In further research [10], while using a common MVPA

978-1-5090-0620-5/16/$31.00

analysis approach and a Neural Network classifier, we were able to accomplish the first task completely. The second and third tasks did not succeed over the cross run and cross subject conditions. (See section 2 for a description of the data.) However, the third task was successful when training and testing was done within one continuous scanning run. (The experimental protocol did not allow this for the second task.) This previous classification scheme consisted of the following processing steps: reading the scans, volume alignment, timing correction (slices alignment), smoothing, Z-Score normalization, feature selection while de-convolving the data and choosing the most 2500 discriminative voxels and classifying the experiment trials using a back propagation feed forward neural network.

The results of both previous works, established that complex linguistic information is decodable from fMRI scans but the need to restrict to the intra-run situation indicated that additional work is needed to compensate for distortions introduced between scanning runs.

In this paper we present a new technique for classifying rapid/fast-event related fMRI experiment trials by generation of artificial data which mimic the measured testing data and investigated if this mechanism can be used for deeper grammatical cognitive tasks. The working hypothesis was that a more complex grammatical task will create a more complex activation in the brain, and make the classification task more difficult at the level of temporal and spatial resolution given by an fMRI signal.

II. Data

Nine subjects were presented with a series of single words and asked to read them aloud. Each of the subjects performed eight distinct runs of 360 seconds each, and was exposed to trials with asterisks, words with and without a root. Four of the eight runs presented only words with diacritic marks and the other four runs presented words only without diacritic marks. The TRs were 2 seconds and the experiment was designed in an event related design. There were baseline shifts between the runs. Each run contained twelve trials with a Hebrew root, twelve trials without a Hebrew root and twelve trials with asterisks. Additional unrelated conditions were presented during each run.

The scans were created in an Analyze format (Analyze, http://www.analyzedirect.com/) and preprocessed: aligned to adjust the movement between volumes (whole image), time corrected to get a signal for the whole brain from the same time point (slices alignment), normalized and smoothed. For full details on the experimental protocol see [3].

III. MVPA classification drawbacks

There are several critical drawbacks in the previous conventional MVPA classification approach.

First, the de-convolved data was used only for choosing the most discriminative voxels at the feature selection procedure. The actual training and testing procedures were done on the raw data before de-convolution. The design of the experiment was fast event related and thus for adjacent trials the evoked BOLD response can overlap in time which can affect the learning and classifying procedures. Thus, it would be preferable to use the trained, de-convolved data on the training procedure in order to reduce the noise that is introduced by the fast event related design.

Second, we need to address the context problem. The prior algorithms populated the classifier with the values of a single TR (the TR on which the trials's HRF response is at its peak while ignoring the initial and post peak undershoots) for each trial, both on training and testing procedures. In this way majority of the temporal information of the full HRF response is lost. Onut and Ghorbani [4] showed that for Block Event-Related fMRI experiment design, it is better to create different classifiers for each voxel that is selected at the feature selection phase and train and test on the full hemodynamic response. Some adjustments are needed to use this approach on Rapid Event-Related design experiments because of the overlapped signal of adjacent trials. To do this we use several adjacent TRs and use them as a single training data point. In other words, each training data point and each testing data point will be the vector that was built from the full response of a specific trial and not only the response at its peak.

IV. The Classification Scheme

A flow-chart of the entire process appears in Figure 1.

A. Conversion to AFNI format

As mentioned, the scans were saved in Analyze format. In event related fMRI experiments designs, the evoked BOLD signal for adjacent stimuli will overlap in time, complicating the identification of signal unique to a specific stimulus.

To handle the convolution of signal and in order to de-convolve it, a conversion to AFNI format [5], [6], [7] was required. AFNI framework offers range of functionality for handling and processing fMRI scans, one of these is de-convolution.

B. Building and training IRF dictionary

This process is illustrated in Fig. 2, sub-figures A - D.

Given the input stimulus function, and the measured fMRI signal data, we need to estimate the impulse response function (IRF) of each voxel for each one of the experiment conditions. By calculating the IRFs we will then be able to build an unbiased, cross run/subject dictionary of IRF estimations which will allow us to build artificial data in the next sections.

We used AFNI's 3dDeconvolve [8, p. 5-20] program to estimate the impulse response function for each experiment condition and every voxel location.

For fMRI data, it is often the case that the measurement can be modeled by:

$$Z_n = y_n + \beta_0 + \beta_1 n + \epsilon_n$$

where β_0 is constant, β_1 is a linear trend coefficient, $\epsilon_n \sim N(0, \sigma^2)$ is uncorrelated Gaussian noise and y_n is the convolution integral of the stimulus condition time course

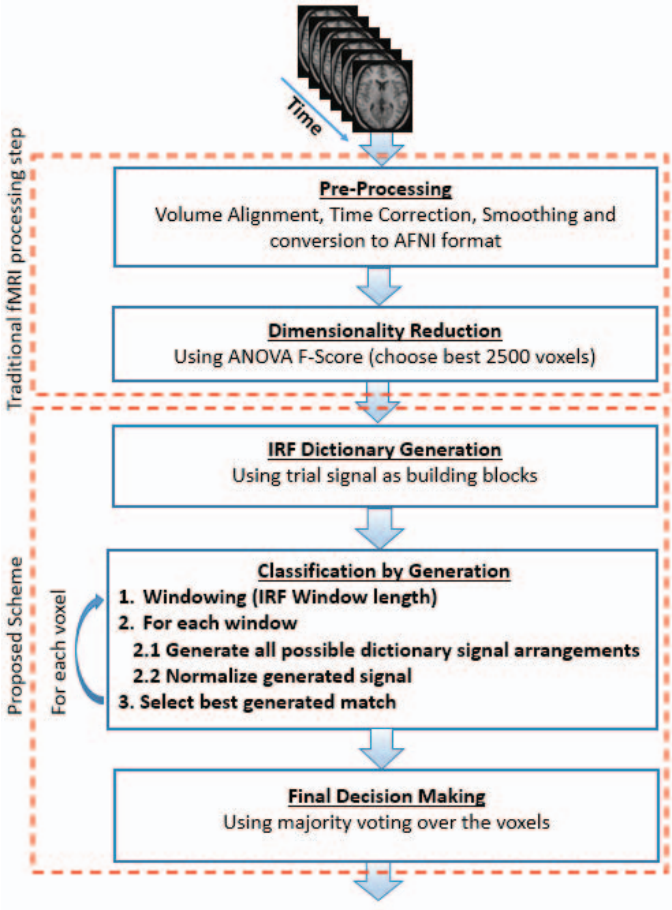

Fig. 1: The Processing Pipeline. Each of the steps in the algorithm is listed in this flowchart. See text for details.

and the approximated hemodynamic response function for the current de-convolved voxel:

$$\begin{aligned} Z_n &= y_n + \beta_0 + \beta_1 n + \epsilon_n \\ &= \beta_0 + \beta_1 n + h_0 f_n + h_1 f_{n-1} \cdots + h_p f_{n-p} + \epsilon_n, \\ for\ n &= p, p+1, ..., N-1 \end{aligned}$$

where $f(t)$ represent the stimulus time course activation function and $h(t)$ the approximated IRF.

3dDeconvolve estimates the impulse response function by performing linear regression to find an estimation vector of

$$\hat{\beta} = \begin{bmatrix} \beta_0 & \beta_1 & h_0 & \dots & h_p \end{bmatrix}$$

which provides the best fit of the training data and the assumed convolved model using minimization of the error sum of squares. This procedure is performed for each voxel separately.

The output impulse response function dictionary stores an unbiased voxel responses, i.e. eliminates baseline shifts in between runs, trend (approximated linear trend) and the approximated Gaussian noise while saving only the averaged response of each voxel for each condition which best fits to the training data. We call these IRFs in the dictionary "atomic" IRFs.

C. Feature Selection

It is well established in machine learning that feature selection can have a substantial effect on the quality of the results (see e.g. [9]). Note that a subset of features can often obtain better classification results due to the fact that not all of the features are essential for the classification process. (Some features can harm the classification process since they just add noise to the system.)

It is also important to note that any feature selection algorithm should take the voxel's convolution into an account.

We followed the common method of producing the most varying features (as produced by ANOVA). This allowed for direct comparison with earlier classification work [10].

We used the GLM (General Linear Model) functionality of AFNI. It generates F-stat values of the IRF dictionary for each voxel which shows how much the voxel's activities varies between conditions over the course of the experiment.

After getting the F-stat map for each voxel we choose the 2500 voxels with the highest F-stat value as the features for the classification.

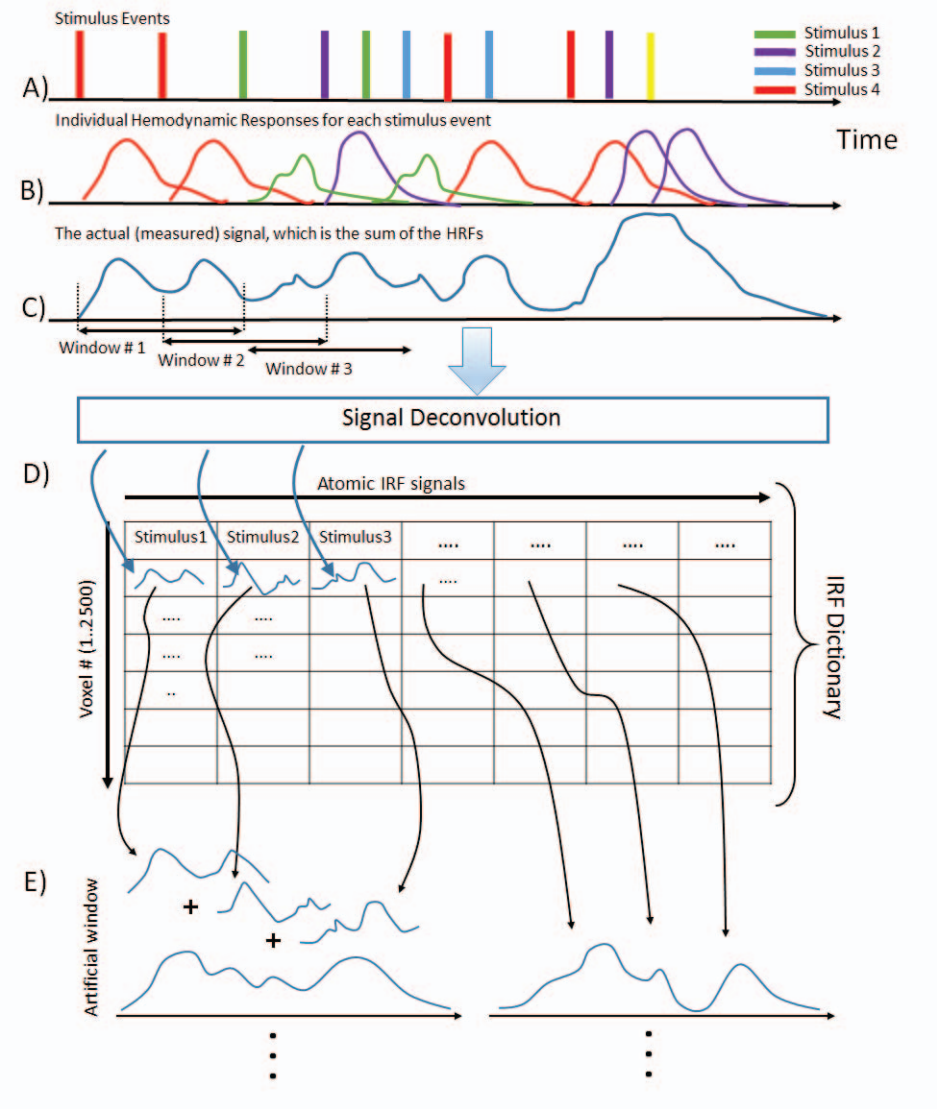

Fig. 2: The Building and Use of the IRF Dictionary. Subfigure A illustrates one run of the timeline of the rapid event experiment. Subfigure B illustrates the IRF signal for each stimulus for a single voxel. Subfigure C illustrates the actual measured signal for this voxel. Subfigure D describe the process of building the table of the atomic IRF signal for each voxel. Subfigure E shows how the atomic signals combine to generate words from which the best-match will be selected. See text for details.

D. Classification

In order to classify the testing set trials we want to be able to classify the whole measured response (and not only the responses of each voxel at the peak of the hemodynamic response function) while taking into account the convolved signal of adjacent trials. Fig. 2, sub-figures D and E illustrate part of this procedure.

1) Data Interpretation: Each testing trial is represented by a windowed activation vector which is a sequence of three dimensional images, i.e. volumes, where each cell in each image represents the voxel's activity. Our activation windows length was set to 10 TRs both in the IRF dictionary and the testing trials. Each such ten element vector window was considered as a single input for the purpose of classification for the event occuring at the start of the window and is assumed to be an activity vector which is a convolution of the several adjacent IRFs produced by events in the window under consideration.

2) Generate Artificial classification windows: To use our trained IRF dictionary and to account for the rapid event related experiment design, we took an observed window from the data and generated artificial windows using the IRF dictionary for all possible combinations of the available conditions occurring at the event times. Each such artificial window includes time shifting of the IRFs to correspond to the

time of the corresponding events. To create each such artificial window we convolved the atomic impulse response functions with the window stimulus time course activation function in the data window to be classified. (Note that since we are working within a single window, in creating the artificial composed IRF we can ignore baseline shifts, linear trend, assumed gaussian noise and amplitude response differences between the conditions.)

We are creating the artificial window for each combination of events and for each voxel separately and thus ending with a $|NumOfFeatures| \times |IRFLength|$ matrix for each combination of events. i.e. the generated artificial word.

3) Normalization: Before using the artificial data to classify the current trial, we first need to normalize both the testing activity windows and the artificial activity windows for each voxel separately and remove any baseline differences between our testing and training windows. These are important steps to perform for the classification procedure. We also normalized the windows to have zero mean and the same Norm value in order us to use cosine similarity correlation test as the classification method.

4) Classification by Correlation: To classify the window, we first compute the correlation between the normalized versions of the measured data windows and the artificial windows of each combination; this is done using cosine similarity. Then, each voxel votes for the first "letter" of the "best match" from the artificial word combinations. The classifier finally choses the "letter" that has the most votes.

E. Test Cases

We tested the classification of each experiment on two resolution levels: run level and subject level.

In order to check the accuracy of the algorithm, we mainly used cross validation tests using the "leave one out" method, taking one observation for testing and keeping all of the other observations for training and building the IRF dictionary.

The tested configuration for each resolution level was:

- Runs Level - for each subject and within each run, creating $N-2$ cross validation tests where on each fold, two trials, one from each condition, left for testing and all the other trials used for training
- Subject Level (Cross Runs) – for each subject, training on three runs and trying to classify the observations on the fourth one, not mixing diacritical runs with non-diacritical runs. Thus, creating $N-1$ cross validation test when $N-1$ runs used for training and one run left for generalization

Note that trying to train over separate runs may introduce some noise as a result of subject movement, head orientation, magnet calibration etc. Thus better results are anticipated if all the work is done over the same person and on the same run.

V. Results

A. Words versus pseudo-words

Table I presents the results. Each class of words, rooted and un-rooted words, was tested against the strings of asterisks. These results are consistent with more standard methods run by Tali Bitan and Yael Weiss using SPM software [2]. Using the same feature selection (ANOVA) as in the previous methods, there is a dramatic increase in the accuracy rate on the intra-run level with mean accuracy of 87.5% with a standard deviation of 8.7%. This task also could be told apart during loud reading at the Subject level but with slightly lower accuracy from the previous performed methods.

Word class	Resolution Level			
	Run Level		Subject Level	
	Classification by Generation	MVPA Analysis	Classification by Generation	MVPA Analysis
Rooted Words	89.1% +/- 8.2%	75.65% +/- 9%	62.7% +/- 10.6%	69.3% +/- 10%
Un-Rooted Words	86.1% +/- 9.4%	75.1% +/- 9%	63.5% +/- 12.9%	70.7% +/- 7%

TABLE I: Words versus pseudo-words

B. Words with Hebrew root versus Words without Hebrew root

Table II presents the results. The results show that using the same feature selection (ANOVA) as in the earlier and more traditional methods, there is a dramatic increase in the accuracy rate for the words with roots versus words without roots task (the most difficult task) on the intra-run level with accuracy of 85.07% and standard deviation of 8.3%. On the cross level using these features, generalization results remained at the chance level. However, using the same feature selection technique (ANOVA) with training and testing over all runs (within subject) there is now a significant level of classification (64% accuracy) for the training set for this task which the previous MVPA analysis could not achieve.

Resolution Level	Words with and without roots	
	Classification by Generation	MVPA Analysis
Run Level	85.07% +/- 8.3%	73.21% +/- 8.9%
Subject Level	50.3% +/- 12.1%	48.97% +/- 8.1%

TABLE II: Words with Hebrew root versus Words without Hebrew root

VI. Conclusions

In this work we tried a method of "best match" between possible responses and data and used this for classification. The motivation was that the standard machine learning method (NN with ANOVA feature selection) could not successfully

classify a deep grammatical task. In fact, the system could not even produce non-random results *even when the testing set was just the training set* in the inter-run situation. The best match over generation is possible in principle because the responses can be represented as a convolution between "basic" responses from a fixed IRF to a given stimulus. This requires three main procedures: (1) finding the possible atomic IRF responses (2) calculating the possible IRF convolved signals from these basic ones (3) making the best match between a data point and one of these responses. In this work, the number of IRF responses is very small and the window on responses can be thought of as fairly short as well. Accordingly the total number of possible IRF signals is also small. In the work presented here, (1) is accomplished by performing de-convolution over all the voxel responses from different fast event related runs into purported responses from corresponding separated events. The appropriate "atomic responses" used in the dictionary is then learned from all of the instances of these responses from the same event. Then (2) is calculated for each voxel by convolving the atomic responses with the appropriate time shifting. (3) is accomplished by a cosine similarity measure on each voxel separately. Finally a direct voting is done by all the voxels to determine the final best match classification from the dictionary. Our results show that this methodology both gave substantially improved results over standard methods in the simpler classification task and in the intra-run case of the complex classification task. Using the ANOVA feature selection, this method does give significant classification when the testing and training set are identical even when the standard methods failed to give significant results for the more complex grammatical task of words with and without roots in the cross-run situation. We emphasize that these results require training data from all runs.

Overall then it seems that the method proposed here, "classification from generation" can give significant results in quite difficult classification tasks, that we were unable to do by earlier, more direct, machine learning techniques. The technique is currently limited to situations (as in most fast event related design protocols) where the number of changed events is relatively small.

ACKNOWLEDGMENT

This research is part of the M.Sc. thesis of Haim Shalelashvili in the Computer Science Department at the University of Haifa. The data was provided by Tali Bitan and Yael Weiss. The computational analysis of the data was performed at the Neuro-Computation Laboratory of the Caesarea Rothschild Institute at the University of Haifa, Israel under the supervision of Prof. Larry Manevitz. Authors are listed in alphabetical order. This work was partially supported by a grant for computational equipment by the Caesarea Rothschild Institute and by a Hardware Grant by NVIDIA Corporation.

REFERENCES

[1] K. A. Norman, S. M. Polyn, G. J. Detre, and J. V. Haxby, "Beyond mind-reading: multi-voxel pattern analysis of fMRI data," Trends in cognitive sciences, vol. 10, no. 9, pp. 424–430, 2006.

[2] W. D. Penny, K. J. Friston, J. T. Ashburner, S. J. Kiebel, and T. E. Nichols, Eds., Statistical Parametric Mapping: The Analysis of Functional Brain Images, 1 edition. Amsterdam , Boston : Academic Press, 2006.

[3] Y. Weiss, T. Katzir, and T. Bitan, "Many ways to read your vowels—Neural processing of diacritics and vowel letters in Hebrew," NeuroImage, vol. 121, pp. 10–19, Nov. 2015.

[4] I.-V. Onut and A. Ghorbani, "Classifying cognitive states from fMRI data using neural networks," in 2004 IEEE International Joint Conference on Neural Networks, 2004. Proceedings, 2004, vol. 4, pp. 2871–2875 vol.4.

[5] R. W. Cox, "AFNI: Software for Analysis and Visualization of Functional Magnetic Resonance Neuroimages," Comput. Biomed. Res., vol. 29, no. 3, pp. 162–173, Jun. 1996.

[6] R. W. Cox and J. S. Hyde, "Software tools for analysis and visualization of FMRI Data," NMR Biomed., vol. 10, pp. 171–178, 1997.

[7] S. Gold, B. Christian, S. Arndt, G. Zeien, T. Cizadlo, D. L. Johnson, M. Flaum, and N. C. Andreasen, "Functional MRI statistical software packages: a comparative analysis," Hum. Brain Mapp., vol. 6, no. 2, pp. 73–84, 1998.

[8] W, B. Douglas. "Deconvolution analysis of fMRI time series data," Milwaukee, WI: Biophysics Research Institute, Medical College of Wisconsin, 2002.

[9] O. Boehm, D. R. Hardoon, and L. M. Manevitz, "Classifying cognitive states of brain activity via one-class neural networks with feature selection by genetic algorithms," Int. J. Mach. Learn. Cybern., vol. 2, no. 3, pp. 125–134, Sep. 2011.

[10] H. Shalelashvili, T. Bitan, A. Frid, H. Hazan, S. Hertz, Y. Weiss, and L. M. Manevitz, "Recognizing deep grammatical information during reading from event related fMRI," in 2014 IEEE 28th Convention of Electrical Electronics Engineers in Israel (IEEEI), 2014, pp. 1–4.