Application of Pattern Recognition Method in a Linguistic Experiment with Unsupervised Classification

Ali Kamel Issmael Junior, Aline Gesualdi Manhães and José Vicente Calvano

***Abstract*— Event-Related Potentials (ERP) are biological electrical signals synchronized with sensory, cognitive or motor stimuli and measured by electroencephalographs (EEG). ERP technique allows non-invasive analysis of brain functions. Based on the results obtained by Soto [1], this work extracts ERP parameters using EEGLAB® and ERPLAB® tools based on Matlab® software [6], [7], [8], [9]. The result of the research was the obtaining of supervised and unsupervised classification scenarios for the classes proposed in the mentioned experiment and the comparative study and discussion of the classification results found, using the methodology proposed by Webb [2]. This article presents the results obtained with unsupervised classification scenarios only and the supervised classification scenarios will be presented in future. The results achieved accuracies very near from the equiprobability, indicating that the use of unsupervised classifiers approaches considered are not adequate to classify Soto's data [1]. This study is innovative in the area of Neurolinguistics, since, at least until now, there are no similar previously published works on the subject found in research databases such as: IEEExplorer; Web of Science; Elesevier and Spring. The results open the possibility of analyzing signals from individuals with this ERP methodology associated to Pattern Recognition, with the possible application of this type of analysis in diagnostic tools, assessment of language learning, among others.**

Keywords*— *ERP, EEG, Pattern Recognition, Linguistics.

I. INTRODUCTION

Event-Related Potentials (ERP) are electrical voltages associated with a neurophysiological response induced by an external event or stimulus. ERPs are obtained by means of Electroencefalography (EEG), which is a non-invasive aparatus sensible enough to measure small electrical potentials in human scalp, as a result of the stimulation by sensory, cognitive or motor events.

This study uses the ERP experimental data of Soto [1] that were addressed underlying cognitive functions of the ERP component measured on target words in sentential and word priming contexts in Portuguese language, for applications in neurolinguistics.

From this ERP experiment data, and the use of specific computer tools EEGLAB® [6] and ERPLAB® [7], [8], based on the software Matlab®[9], is possible to treat these experimental data to investigate if there are specific parameters of recognition for the ERP signals related to each kind of stimuli. The treatment of these ERP signals involves the study of digital signal processing techniques applied with pattern recognition theory.

The goal of this work is to investigate, by applying the pattern recognition methodology proposed by Webb [2], on the ERP results from the Soto [1] data experiment, if it is possible to obtain good classification scenarios. Considering each type of stimulus for the epochs not labeled (unsupervised classification and clustering methods), this work verifies the consistency of the sentences and words classes proposed by Soto [1], through the extraction of attributes of the EEG and ERP signals.

This study is innovative in the area of Neuroinguistics, since, at least until now, there are no similar previously published works on the subject found in research databases such as: IEEExplorer; Web of Science; Elesevier and Spring. The results open the possibility of analyzing signals from individuals with this ERP methodology associated to Pattern Recognition, with the possible application of this type of analysis in diagnostic tools, assessment of language learning, among others.

A. EEG and ERP theory

The ERPs uses an EEG procedure that measures electrical activity of the brain over time using electrodes placed on the scalp. The EEG reflects thousands of simultaneously ongoing brain processes in specific points distributed in specific areas and Regions of Interest (ROI) of the scalp depending of the cognitive target of the research. These ROI are related more with the cognitive analysis process than the data collection that is related with the individual electrodes. The brain response to a single stimulus or event of interest is not usually visible in the EEG recording of a single trial. The ERP technique consists on a signal amplification that adds up and averages specifically time-locked epochs, which are replications of a stimulus, and ideally can present a lower signal-to-noise ratio (SN) than those of the original waveforms. The signal to noise ratio (SN) is a quality measure to signal processing. SN is the ratio between the signal power and the noise power [3].

The EEG signal is recorded as a continuous signal, and stimulus presentation is marked for onset, and is difficult to be detected and marked. The raw signal is usually filtered for low frequencies (e.g. high pass of 0.01Hz) and amplified. After that, a computer or a separately connected trigger box marks a digital code or a pulse width on the recorded signal, allowing the marking on the continuous EEG signal of the exact onset of the stimulus, and type of stimulus shown. An illustrative example of an experiment is presented schematically in Figure 1: subjects saw many "X"s, sparsely alternated by "O"s. The fragments, called epochs, related to the event are averaged for each electrode so as to amplify the response and filter out noise

Ali Kamel Issmael Junior¸ M.Sc. Student of Department of Electrical Engineering, Federal Center of Technological Education Ceslso Suckow da Fonseca (CEFET-RJ),, Rio de Janeiro-RJ, Brazil, E-mails: alikamel@ig.com.br

Aline Gesualdi Manhaãs¸ D.Sc., Department of Electrical Engineering, Federal Center of Technological Education Ceslso Suckow da Fonseca (CEFET-RJ), Rio de Janeiro-RJ, Brazil, E-mails: alinegesualdi@gmail.com.

José Vicente Calvano¸ D.Sc., Naval War School (EGN), Marinha do Brasil, Rio de Janeiro-RJ, Brazil, E-mails: jvcalvano@gmail.com.

coming from other neurophysiologic activity or interference of electrical equipment. These averaged responses, the ERPs, can now be compared and characterized in terms of amplitude (in μV) - the peak of the wave - and latency (in ms) - the time in which the wave peaks [1]. The figure 1 shows a simplified schema for ERP experiment.

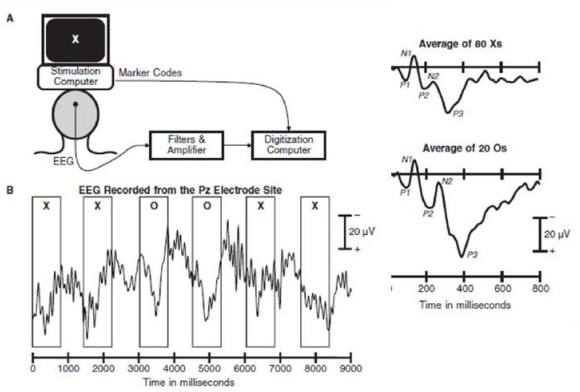

Fig. 1. Example of EEG/ERP experiment with epochs "X" and "O" [4]

ERP signals resulted from the experiment present a series of positive and negative voltage deflections, which are related to a set of underlying components called ERP components. The usual ERP components are referred to by a letter (N/P) indicating polarity (negative/positive), followed by a number indicating either the latency in milliseconds or the component's ordinal position in the waveform [4].

For example, a negative-going peak that is the first substantial peak in the waveform and often occurs about 100 milliseconds after a stimulus is presented is often called the N100 (indicating its latency is 100 ms after the stimulus and that it is negative) or N1 (indicating that it is the first peak and is negative); it is often followed by a positive peak, usually called the P200 or P2. The stated latencies for ERP components are often quite variable. For example, the P300 component may exhibit a peak anywhere between 250ms - 700ms [5]. In the Figure 2 below, an example of an ERP waveform signal with these components can be seen.

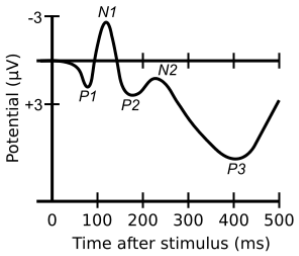

Fig. 2. A fictitious illustrative waveform graph example showing several ERP components as P1 (P100), N1(N100), P2(P200), N2 (N200) and P3 (P300) [4]

As described by Woodman [5], an ERP component can be simply defined as one of the component waves of the more complex ERP waveform. ERP components are defined by their polarity (positive or negative going voltage), timing, scalp distribution, and sensitivity to task manipulations. Different ERP component nomenclatures emphasize different aspects of these defining features and to provide a jumping off point for literature reviews.

Concerning the target of this work (language specific properties), Soto [1] indicates that indeed ERP methodologies have brought much evidence to show that very detailed linguistic information has an immediate effect on processing streams. Soto [1] also indicate that the N400 component of the ERP signal can be influenced by strict linguistic variables.

For this study, three ERP parameters that were extracted from the experiment. These parameters were:

a) Mean Amplitude Between two fixed latencies - Mean Peak Amplitude in an ERP Time Range;

b) Peak Amplitude - the maximum peak amplitude in a ERP Time Range, and

c) Peak Latency - the time value for the oiccurence of the maximum peak amplitude.

B. Soto [1] ERP Experiment.

In order to investigate the specific nature of the N400 effects in sentence and word pair contexts, Soto [1] proposed a sentence and word priming tasks in Portuguese language. Following from the proposed variables [1], 4 conditions and one control condition were established for the sentence task: (i) congruous supportive-context (CSC): e.g. "Até sem capacete, João dirige a moto feito louco" ("Even without a helmet, João drives the bike like a crazy"); (ii) congruous non-supportive context (CNSC), e.g. "Todos os dias, João dirige a moto feito louco" ("Every day, John drives the bike like a crazy"); (iii) incongruous supportive-context (ISC): e.g. "Até sem capacete, João dirige a pera feito louco" ("Even without a helmet, João drives the pear like a crazy"); and (iv) incongruous non-supportive context (INSC), e.g. "Todos os dias, João dirige a pera feito louco" ("Every day, John drives the pear like a crazy"). For the word pair task, Soto [1] proposed 3 conditions and one control condition were established: (i) associative semantic relation (ASR): e.g."ÔNIBUS moto" ("BUS motorbike"); (ii) syntactic and semantic relation (SSR): e.g. "CAPACETE moto" ("HELMET motorbike"); (iii) unrelated pair (UR) "FACA nuvem" ("KNIFE cloud") ; and control 2: a pair with pseudo word (PW) target: e.g. "CARRO garufa" ("CAR garufa").

Concerning the experimental setup, 21 university students participated in the study (female =11), distributed evenly over 4 versions, average age 22 years old, all right-handed, with normal or corrected-to-normal vision Participants' judgments were recorded by pressing with one of two fingers of the right hand either a red or a green button on a button box. The position of the green and red buttons, destined for YES and NO responses, was swapped for each participant. The Figure 3 illustrate the experiment:

Fig. 3. Soto´s experiment [1]

The scalp Regions of Interest (ROIs) are presented in the Figure 4. The ROIs along the mid-line were: Frontal (F1-ch34, F2-ch60, FC1-ch7, FC2-ch27, FCz-ch38 and Fz-ch2); Central (C1-ch39, C2-ch56, CP1-ch11 CP2-ch21, CPz-ch52 and Cz-

ch22), Parietal (CP1-ch11, CP2-ch21, CPz-ch52, P1-ch43, P2-ch51, and Pz-ch12), and Occipital (O1-ch15, O2-ch17, Oz-ch16, PO3-ch46, PO4-ch48, and POz-ch47). On the left hemisphere, they were Frontal (F3-ch3, F5-ch35, F7-ch4, FC3-37, FC5-ch6 and FT7-ch36); Central (C3-ch8, C5-ch40, CP3-ch42, CP5-ch10, T7-ch9 and TP7-ch41), Parietal (CP3-ch42, CP5-ch10, P3-ch13, P5-ch40, P7-14 and TP7-ch41), and Occipital (P3-ch13, P5-ch44, P7-ch14, PO3-ch46 and PO7-ch45). And on the right hemisphere, they were: Frontal (F4-ch28, F6-ch59, F8-ch29, FC4-ch57, FC6-ch26 and FT8-ch58); Central (C4-ch23, C6-ch55, CP4-ch53, CP6-ch20, T8-ch24 and TP8-ch54), Parietal (CP4-ch53, CP6-ch20, P4-ch18, P6-ch50, P8-ch19 and TP8-ch54), and Occipital (P4-ch18, P6-ch50, P8-ch19, PO4-ch48 and PO8-ch49).

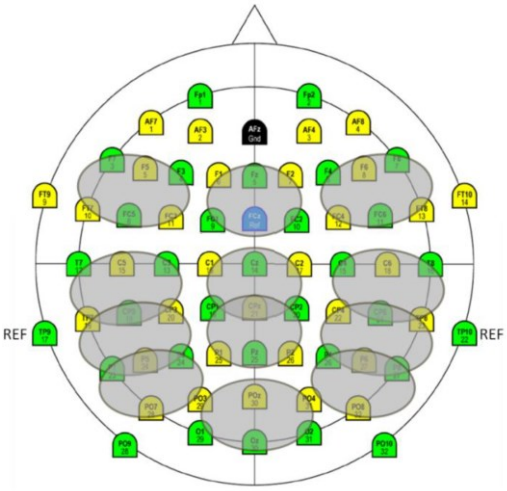

Fig. 4. ROI definition as based on anatomical proximity [1]

To obtain the ERP signals for each ROI, it is necessary to add the contribution of each electrode channel related with the Region and take the arithmetic media. So, considering the electrode distribution of the experiment, the ERP signal for each ROI is obtained by the following equations:

Frontal Mid Line (ch63) = (ch2 + ch7 + ch27 + ch34 + ch38+ ch60)/6 (1)

Central Mid Line (ch64) = (ch11 + ch21 + ch22 + ch39 + ch52 + ch56)/6 (2)

Pariental Mid Line (ch65) = (ch11 + ch12 + ch21 + ch43 + ch51 + ch52)/6 (3)

Occipital Mid Line (ch66) = (ch15 + ch16 + ch17 + ch46 + ch47 + ch48)/6 (4)

Frontal Left Side (ch67) = (ch3 + ch4 + ch6 + ch35 + ch36 + ch37)/6 (5)

Central Left Side (ch68) = (ch8 + ch9 + ch10 + ch40 + ch41 + ch42)/6 (6)

Pariental Left Side (ch69) = (ch10 + ch13 + ch14 + ch41 + ch42 + ch44)/6 (7)

Occipital Left Side (ch70) = (ch13 + ch14 + ch44 + ch45 + ch46)/5 (8)

Frontal Right Side (ch71) = (ch26 + ch28 + ch29 + ch57 + ch58 + ch59)/6 (9)

Central Right Side (ch72) = (ch20 + ch23 + ch24 + ch53 + ch54 + ch55)/6 (10)

Pariental Right Side (ch73) = (ch18 + ch19 + ch20 + ch50 + ch53 + ch54)/6 (11)

Occipital Right Side (ch74) = (ch18 + ch19 + ch48 + ch49 + ch50)/5 (12)

C. *EEG/ERP Data Software toolboxes and Matlab® platform*

To extract and organize the ERP data from an experiment there are several softwares that help in this data mining activity. Concerning this work, EEGLAB® and the ERPLAB®, which are Matlab® toolboxes for processing and analyzing EEG and ERP data were used, and for the digital process and pattern recognition study, Matlab® was used.

D. *Pattern Recognition Theory*

Pattern recognition systems are in many cases trained from labeled "training" data (supervised learning or discrimination), but when no labeled data are available other algorithms can be used to discover previously unknown patterns (unsupervised learning or clustering). Webb [2] defines that in supervised classification a set of data samples (each consisting of measurements on a set of variables or attributes or features that can be extracted) are associated with which correspond to the class types. These classes and features are used in the classifier design. In unsupervised classification, the data labels (classes) are not known and it is necessary to seek for groups in the data with the same characteristics, by the features that can distinguish one group (class) from another.

As described by Webb [2], an oversimplified procedure of pattern recognition is shown in the Figure 5 with the roles of all data experiment origin and software tools used in this work.

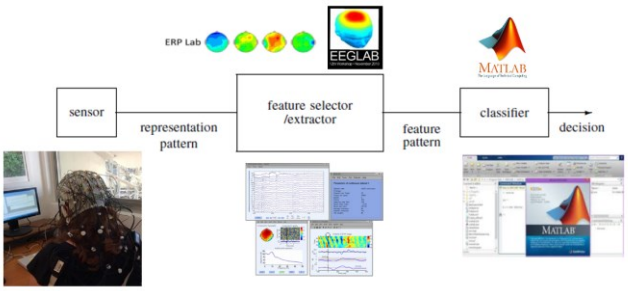

Fig. 5. Pattern Recognition Method and roles of each software tool used [2]

As shown in the Figure 5, in this study, the Soto [1] experiment is related with the Sensor block (in fact, the EEG/ERP experiment) and the feature selector/extractor is related with the softwares EEGLAB® and ERPLAB®. The classifier box was tailored with the software Matlab®.

II. METHODOLOGY

The methodology considered in this work is based in the stages in a pattern recognition problem indicated by Webb [2]. The Figure 5 shown this procedure in a flowchart way:

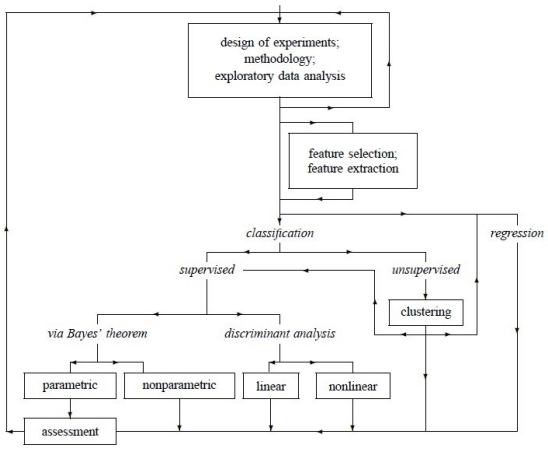

Fig. 6. Webb´s Pattern Recognition methodology [2]

In fact, the methodology adopted use the following adapted steps:

1. Formulation of the problem, Data collection and Initial examination of the data;

2. Feature selection or feature extraction;

3. Unsupervised pattern classification or clustering;

4. Supervised pattern classification;

5. Assessment of results and Interpretation

As mentioned in the Abstract, this article will focus only in the unsupervised pattern classification results, not addressing the supervised classifiers results obtained.

In this study from Soto [1] experiment, the features that were extracted for the words and sentences task are the ERP parameters Mean Amplitude Between two fixed latencies, Peak Amplitude, Peak Latency, in addition to the ERP Time Range, the Region of Interest (ROI) and the human subject related to each measurement. Concerning the classes, for the sentences task, they are S1 (CSC), S2 (CNSC), S3 (ISC), S4 (INSC) and S5 (Control) and, for the words task, they are S1 (SSR), S2 (ASR), S3 (Control 1 - UR) and S4 (Control 2 - PW).

The parameters/features Mean Amplitude Between two fixed latencies, Peak Amplitude, Peak Latency with their respectively values to use in the algorithms were extracted from the software EEGLAB® and ERPLAB®.

For classification purposes, before building the classifier, it is necessary to use numeric single values for the features and classes, to allow the convergence of the classification methods of Matlab®. Features or classes that are not numerical variables, for instance, the Region of Interest (ROI) feature, shall be coded with numerical values with a coherent correspondence with the original string value.

In the Tables I and II below, the data organization and coding of values for the other features and classes for the elaboration of the classifiers are shown.

TABLE I. WORDS AND SENTENCES TASK ORGANIZATION AND CODING FOR FEATURES

Feature	Real Value	Code for Matlab® algorithm
ERP Time Range	150-300ms	1
	300-500ms	2
	500-700ms	3
Region of Interest (ROI)	Frontal Mid Line	1
	Central Md Line	2
	Pariental Mid Line	3
	Occiptal Mid Line	4
	Frontal Left Side	5
	Central Left Side	6
	Pariental Left Side	7
	Occiptal Left Side	8
	Frontal Right Side	9
	Central Right Side	10
	Pariental Right Side	11
	Occiptal Right Side	12
Subject	2	2
	3	3
	4	4
	5	5
	6	6
	7	7
	8	8
	9	9
	10	10
	13	13
	15	15
	16	16
	17	17
	18	18
	19	19
	20	20

TABLE II. WORDS AND SENTENCES TASK ORGANIZATION AND CODING FOR CLASSES

Task	Classes	Code for Matlab® algorithm
Words	S1 (SSR)	1
	S2 (ASR)	2
	S3 (Control 1 - UR)	3
	S4 (Control 2 - PW)	4
Sentences	S1 (CSC)	1
	S2 (CNSC)	2
	S3 (ISC)	3
	S4 (INSC)	4
	S5 (Control)	5

After the use of the EEGLAB® and ERPLAB®, the following input data for the classifiers was extracted:

a) For Sentences Task: a matrix with 2880 lines and 7 columns, where the features corresponding to: column A: Mean Amplitude Between two fixed latencies, column B: Peak Amplitude, column C: Peak Latency, column D: Region of Interest (ROI), column E: ERP Time Range and column F: Subject index. The last column G corresponding to the classes.

b) For Words Task: a matrix with 2304 lines and 7 columns, where the features corresponding to: column A: Mean Amplitude Between two fixed latencies, column B: Peak Amplitude, column C: Peak Latency, column D: Region of Interest (ROI), column E: ERP Time Range and column F: Subject index. The last column G corresponding to the classes.

The Figure 7 shows this input data organization.

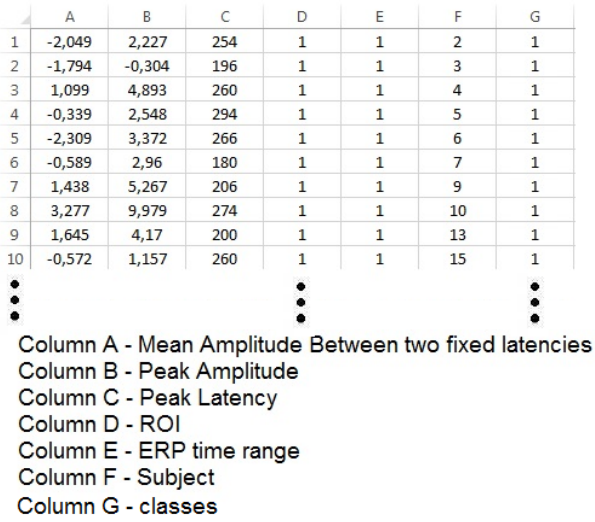

Fig. 7. Input data organization extracted from the EEGLAB® and ERPLAB® for the classifiers.

This study considered for the unsupervised classification and clustering all data for the classifiers in one single dataset, not dividing in subsets as training, validation or test. As the classifiers are unsupervised, the column G related to the labels for the classes are not used in the classification. The Matlab® clustering methods used to create the classifiers algorithms scripts are Hierarchical Clustering, k-means and Gaussian Mixture Models (GMM). The GMM method allows to use only 2 features in the clustering. Because this characteristic, the features Mean Amplitude Between two fixed latencies, Peak Amplitude and Peak Latency were combined two-by-two to build the classifiers. The figure of merit used was the Classification Accuracy, that is the number of correct predictions from all predictions made.

III. RESULTS AND DISCUSSION

The results will be shown with the best Total Accuracy achieved for each classifier. For the unsupervised classifiers, the best results for Sentences Task are presented in the Table III.

TABLE III. UNSUPERVISED CLASSIFIERS RESULTS FOR SENTENCES TASK

Classifier	Accuracies for Sentences Task	Observation concerning the Parameters used
Hierarchical Clustering	21,63 %	"pdist" metric "cityblock" with a "linkage" method "average" "pdist" metric "cityblock" with a "linkage" method "centroid"
K-means	52,92 %	2 clusters with k-means metric "cityblock"
	19,44 %	5 clusters with k-means metric "cityblock"
	53,33 %	2 clusters with k-means metric "sqEuclidean"
	18,13 %	5 clusters with k-means metric "sqEuclidean"
Gaussian Mixture Models	19,24 %	Features used: Mean Amplitude Between two fixed latencies and Peak Amplitude
	21,18 %	Features used: Mean Amplitude Between two fixed latencies and Peak Latency
	19,24 %	Features used: Peak Amplitude and Peak Latency

For the unsupervised classifiers, the best results for Words Task are presented in the Table IV.

TABLE IV. UNSUPERVISED CLASSIFIERS RESULTS FOR WORDS TASK

Classifier	Accuracies for Sentences Task	Observation concerning the Parameters used
Hierarchical Clustering	28,21%	"pdist" metric "spearman" with a "linkage" method "single"
K-means	48,44 %	2 clusters with k-means metric "cityblock"
	23,26 %	4 clusters with k-means metric "cityblock"
	32,68 %	3 clusters with k-means metric "sqEuclidean"
	25,00 %	4 clusters with k-means metric "sqEuclidean"
Gaussian Mixture Models	24,91 %	Features used: Mean Amplitude Between two fixed latencies and Peak Amplitude
	24,78 %	Features used: Mean Amplitude Between two fixed latencies and Peak Latency
	24,39 %	Features used: Peak Amplitude and Peak Latency

As the results indicated, the accuracies results achieved are very near from the equiprobability. This means that the clustering and unsupervised classification methods considered are not appropriate for the classification task for Soto [1] experiment.

IV. CONCLUSIONS

The objective of this work that was to investigate Webb's pattern recognition methodology [2] in ERP results from Soto [1] data experiment to classify correctly different patterns was considered as achieved wherever the results obtained with the clustering and unsupervised classification.

EEGLAB®, ERPLAB® and Matlab® were used as software tools to perform pre-processing and pattern recognition steps in Soto's EEG data and worked properly. As shown in this article, clustering and unsupervised classification methods used were not appropriated for the classification task.

Other methods of classification for clustering and unsupervised classification shall be considered in order to study different aspects of this dataset, not only for pattern recognition, but also for their use as a possible method in neurolinguistics and medicine through the identification of which ERP features can be more influent in the classification process.

The supervised classifiers developed and their results will be discussed later in another article.

ACKNOWLEDGEMENTS

The authors would like to thanks Marije Soto and the Federal Center of Technology Celso Suckow da Fonseca to allow the execution of this work.

REFERENCES

[1] SOTO, Marije*, ERP and fMRI Evidence of Compositional Differences between Linguistic Computations for Words and Sentences. Marije Soto - Rio de Janeiro:UFRJ./Faculdade de Letras, 2014.*

[2] WEBB, Andrew R., Statistical Pattern Recognition, Second Edition. John Wiley & Sons, Ltd. 2012. ISBNs: 0-470-84513-9 (HB); 0-470-84514-7 (PB)

[3] GESUALDI, Aline da Rocha; FRANÇA, Aniela Improta. Event-related brain potentials (ERP): an overview. Revista Linguística / Revista do Programa de Pós-Graduação em Linguística da Universidade Federal do Rio de Janeiro. Volume 7, número 2, dezembro de 2011. ISSN 1808-835X 1. [http://www.letras.ufrj.br/poslinguistica/revistalinguistica]

[4] LUCK, Steven J., An Introduction to the Event-Related Potential Technique, Massachusetts Institute of Technology MIT Press books, 2nd Ed., 2014, ISBN 978-0-262-52585-5

[5] WOODMAN, Geoffrey F.. A Brief Introduction to the Use of Event-Related Potentials (ERPs) in Studies of Perception and Attention. Atten Percept Psychophys. 2010 November; 72(8): doi:10.3758/APP.72.8.2031.

[6] EEGLAB®. EEGLAB® Tutorial. Available on: <http://sccn.ucsd.edu/wiki/Getting_Started>, accessed in April, 15th, 2016.

[7] ERPLAB®. ERPInfo-ERPLAB® Toolbox. Available on:<http://www.erpinfo.org/erplab.html>, consulted on April, 15th, 2016.

[8] LOPEZ-CALDERÓN, Javier, LUCK, Styeven J. ERPLAB: an open-source toolbox for the analysis of event-related potentials. Frontiers in Human Neuroscience. Volume 8. Article 213. p. 2-14. April 2014.

[9] MATLAB®. MATLAB - The Language of Technical Computing. Available on: <https://www.mathworks.com/products/matlab.html>, consulted on April, 15th, 2016a.