Towards a Semantic Calibration of Lexical Word via EEG

Marios Poulos

Laboratory of Information Technology, Departement of Archive and Library Science, Ionian University, Ioannou Theotoki 72, 49100 Corfu, Greece
mpoulos@ionio.gr

Abstract. The calibration method used in this study allows for the examination of distributed, but potentially subtle, representations of semantic information between mechanistic encoding of the language and the EEG. In particular, a horizontal connection between two basic Fundamental Operations (Semantic Composition and Synchronization) is attempted. The experimental results gave significant differences, which can be considered reliable and promising for further investigation. The experiments gave helpful results. Consequently, this method will be tested along with the classification step by appropriate neural network classifiers.

Keywords: Hybrid method, EEG, Signal Processing, Semantic Processing, Lexical Word

1 Introduction

With the blossoming of the Internet, the semantic interpretation of the word is now more imperative than ever. Two scientific approaches can lead one to achieve this aim: linguistic formalism and the neuroscience method.

Linguistic formalisms are served by semantic nets (such as ontology schema), and providing by a well-defined semantic syntax, which are also combining features of object-oriented systems, of frame-based systems, and of modal logics. However, the use of these systems creates many problems. The main problem of information extraction systems is low degree of portability due to language dependent linguistic resources and to domain-specific knowledge (ontology) [1]. Additionally, the individual differences in information needs, polysemy (multiple meanings of the same word), and synonymy (multiple words with same meaning) pose problems [2] in that a user may have to go through many irrelevant results or try several queries before finding the desired information. Although, using ontologies to support information retrieval and text document processing has lately involved more and more attention, existing ontology-based methods have not shown benefits over the outdated keywords-based Latent Semantic Indexing (LSI) technique [3]. A partial solution to the above problems uses the semantic measurement of similarity between words and terms, which plays an important role in information retrieval and information integration [4, 5]. Nowadays, this measurement is implemented by the allocation of the words in a metric space, which is called semantic map [6]. Many methods have

been developed for this aim such as maps based on the representation of semantic difference of the word as geometrical distance [7–10] and the maps that depict the semantic positions of the words using the likelihood of the word which appears in a particular topic or document [6].

In neuroscience practice, the problem of data sharing in brain electromagnetic research, similar to other scientific fields, is challenged by data scale, multivariate parameterizations, and dimensionality [11]. The research about organization and localization of lexico-semantic information in the brain has been discussed in the past. Decoding methods, on the other hand, allow for a powerful multivariate analysis of multichannel neural data. A significant work about this problem showed the decoding analysis to demonstrate that the representations of words and semantic category are highly distributed both spatially and temporally [12]. In particular, many studies in the past showed that the suitable Support Vector Machines (SVMs) [14,13], which have been constructed by decoding multichannel EEG data, possess critical features in relation to the conceptual understanding of written words. These features are depicted in an acceptable time span of a 300-600 mc EEG recording and especially in spectral features (8–12 Hz) power [14, 12]. Furthermore, in recent work [15] the EEG decoding of semantic category reveals distributed representations for single concepts is implemented by applying data mining and machine learning techniques to single trials of recorded EEG signals.

However, until now, the gap between the linguistics and the neuroscience has been considered unbridgeable [16]. This is illustrated in Table 1.

Table 1. The two unordered lists enumerate some concepts canonically used to explain neurobiological or linguistic phenomena. There are principled ontology-process relationships within each domain (i.e., vertical connections) [16]

]**Linguistics**	**Neuroscience**
Fundamental Elements	
Distinctive Feature	Dendrites, spines
Syllable	Neuron
Morpheme	Cell-Assembly/Ensemble
Noun Phrase	Population
Clause	Cortical Column
Fundamental Operations on Primitives	
Concatenation	Long-Term Potentiation
Linearization	Receipt Field
Phrase-Structure	Oscillation
Semantic Composition	Synchronization

This study attempts to bridge the gap between the two methodologies. In particular, a horizontal connection between two basic Fundamental Operations (Semantic Composition and Synchronization) is proposed via a Semantic Calibration of Lexical Word via EEG. The idea is based on the following four (4) approaches:

1. The determination of any ordered sequence of k characters occurring in each word. This approach follows the Kernel learning Philosophy [17,18] and consists of an early semantic interpretation of the word "on step beyond of the word" [19]
2. The isolation of a significant feature of an EEG segment 500ms duration according to aforementioned reference is attempted[16, 14, 12]
3. A new signal generation is derived from ordered sequence of k characters and the suitable modulated EEG signal.
4. Features are extracted from the new signal and statistical testing of the semantic feature.

2 Method

The section is divided into four subsections. In the first subsection, "Numerical Encoding of Word's characters," the determination of any ordered sequence of k characters occurring in each word is considered. Preprocessing of the EEG signal and feature extraction is described in the second subsection. Data acquisition, to be used in the experimental part, is outlined in the third subsection. And the fourth subsection presents a statistical approach to the semantic feature of this calibration.

2.1 Numerical Encoding of Word's characters

At this stage, the characters of the selected word are considered as input vector. Then, using a conversion procedure where a symbolic expression (in our case an array of characters of a word) is converted to ASCII characters in a string of arithmetic values. As a result, we obtained a numerical value vector for each. These values ranged between 1–128.

Thus, a vector $\vec{\mathbf{a}}$ with length k is constructed, where k is the number of characters in each investigated word.

2.2 Preprocessing of the EEG signal

Electroencephalographic (EEG) data contains changes in neuro-electrical rhythm measured over time (on a millisecond timescale), across two or more locations, using noninvasive sensors ("electrodes") that are placed on the scalp surface. The resulting measures are characterized by a sequence of positive and negative defections across time at each sensor. For example, to examine brain activity related to language processing, the EEG may be recorded during donation of the words, using 128 sensors in a time span of 500ms. In principle, activity that is not event-related will tend toward zero as the number of averaged trials increases. In this way, ERPs provide increased signal-to-noise (SNR) and thus increased sensitivity to functional (e.g., task-related) manipulations [11].

In order to model the linear component of an EEG signal $x(n)$ known to represent the major part of its power (especially in the alpha rhythm frequency band), the selected segment is submitted in alpha rhythm filtering. As it is known, the alpha

rhythm is the spectral band of 8-12 Hz, extracted from the original EEG spectrum and recorded mainly from the occipital part of the brain, when the subjects are at rest with their eyes closed. Thus, the spectral values of the EEG signal are obtained and then restricted to the alpha rhythm band values only in a new signal $y(n)$ which becomes from the time domain difference equation describing the general Mth-order IIR filter, having N feed forward stages and M feedback stages in filter cut upper (a) and lower (b) limit. The time domain expression for an Mth-order IIR filter is given by the following equation (1):

$$y(n) = b(0)x(n) + b(1)x(n-1) + b(2)x(n-2) + \ldots + b(N)x(n-N) + a(1)y(n-1) + a(2)y(n-1) + \ldots a(M)y(n-M) \quad (1)$$

2.3 Semantic Calibration of Lexical Words via EEG

In this stage, in order to create a new signal $z(n)$ with specific hybrid features, the vector

$$\mathbf{a} = \begin{bmatrix} a_1 \\ . \\ . \\ . \\ a_k \end{bmatrix} \quad (2)$$

and the signal $y(n)$ are combined by the following steps:

1. The length of the signal $y(n)$ is divided in k equal segments, where each has length $l = \| \frac{n}{k} \|$ and is given by the following equation (3)

$$y(n) = \sum_{n=1+(i-1)l}^{il} y(n), \text{ where } i = 1, \ldots, k \quad (3)$$

2. A new signal $z(n)$ is generated by the residuals between the vector $\mathbf{a}$ and signal $y(n)$. The calculation takes place for each character per segment and is depicted in the following equation:

$$z(n) = \sum_{n=1+il}^{(i+1)l} \left(y(n) - a_i \right) \quad (4)$$

2.4 The AR model- Feature Extraction

The linear component of the signal $z(n)$ is implemented via a linear, rational model of the autoregressive type, AR [20]. This signal is treated as a superposition of a signal component (deterministic) plus additive noise (random). Noise is mainly due to imperfections in the recording process. This model can be written as

$$x_t + \sum_{i=1}^{p} b_i x_{t-1} = 0 \tag{5}$$

It is an independent, identically distributed driving noise process with zero mean and unknown variance σ_e^2 ; model parameters $\{b_i \quad, i = 1, 2,..., p\}$ are unknown constants with respect to time.

It should be noted that the assumption of time invariance for the model of the text vector can be satisfied by restricting the signal basis of the method to a signal "window" or "horizon" of appropriate length.

The linear model can usually serve as a (more or less successful) approximation when dealing with real world data. In the light of this understanding, the linear model is the simpler among other candidate models in terms of computing spectra, covariances, etc.

In this work, a linear model of the specific form AR(p) is adopted. The choice of the order of the linear models is usually based on information theory criteria such as the Akaike Information Criterion (AIC) [21], which is given by

$$AIC(r) = (N - M)\log\sigma_e^2 + 2r \tag{6}$$

where,

$$\sigma_e^2 = \frac{1}{N - M}\sum_{t=M+1}^{N} e_t^2 \tag{7}$$

N represents the length of the data record; M is the maximal order employed in the model; (N-M) is the number of data samples used for calculating the likelihood function; and r denotes the number of independent parameters present in the model. The optimal order r* is the minimizer of AIC(r).

We have used the AIC to determine the order of the linear part of the model in i.e. the optimal order p of the AR part of the model. For each candidate order p in a range of values [pmin, pmax], the AIC(p) was computed from the residuals of each record in the ensemble of the EEG records available. This is because we deal with recordings of real world data rather than the output of an ideal linear model. We have thus seen that AIC(p) takes on its minimum values for model orders p ranging between 5 and 8, record-dependent. In view of these findings, we have set the model order of the AR part to p = 7, for parsimony purposes [22].

2.3 Identification Procedure

In this stage, the extracted sets of the 7 order AR coefficients x of the generated signal Z(n) are submitted to compute the difference between the variances for two response variables—see equation (8).

$$s = \frac{\sum_{i=1}^{n} (x - \bar{x})^2}{n-1} \tag{8}$$

For the difference of the variances, the variance is computed for each of the two samples before their difference is taken.

3 Experimental Part

As example of this study, the same simple words are used with related study [18] in order to the degree of contiguity between homonymous words to be investigated. For this reason, the words "cat," "car," "bat," "bar" are investigated, and all the algorithms according to the aforementioned method are applied. More details are depicted in figures 1 and 2. Thereafter, the AR coefficients are extracted for each word (an example of this is presented on Table 2). Finally, in Table 3 the differences between each of the pairs of AR coefficients are isolated. Specifically, the isolation of a significant feature of an EEG segment 500ms duration according to aforementioned reference and the appropriate filtering is depicted in fig 1
The calibration of the characters of each word on the new filtering signal is presented in figure 2
The determination of any ordered sequence of k characters occurring in each word is depicted in table 2 as well as the difference between the variances of the tested words which are presented in table 3.

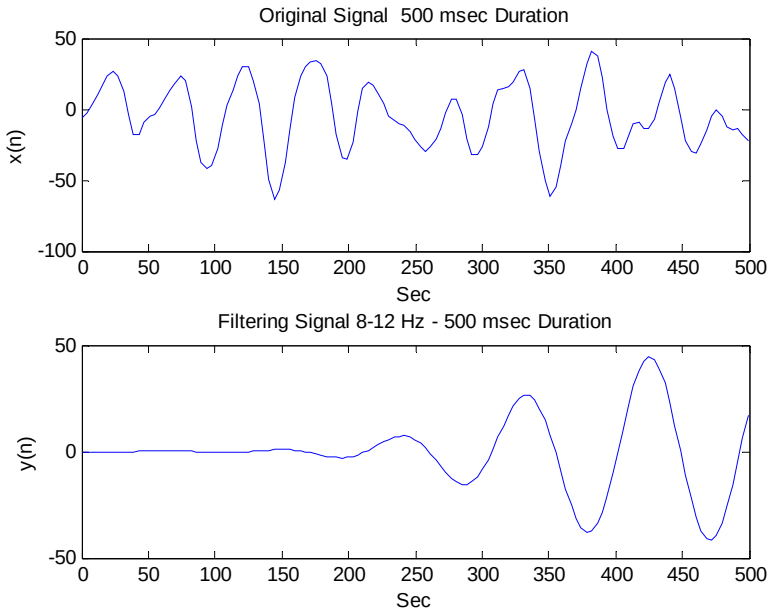

Fig. 1. The first two steps (EEG Selection and Filtering) of the proposed method are applied

Table 2. The AR coefficients of the word "car" from the generated signal z(n).

AR coefficients "car"
0.8008
-0.2886
0.0460
0.2848
0.2151
-0.0251
-0.2828
0.8008

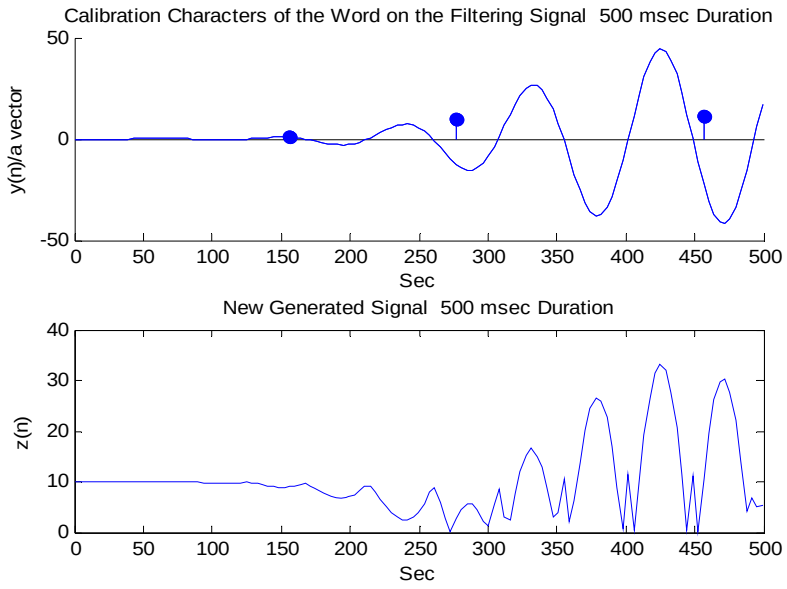

Fig. 2. The calibration of the word "car" on the filtered signal y(n) is presented in the upper figure, while in the below figure the generated signal z(n) is depicted.

Table 2. The two unordered lists enumerate some concepts canonically used to explain neurobiological or linguistic phenomena. Principled ontology-process relationships connect words in each domain (i.e., vertical connections) [16]

Identification Procedure (Difference of Variance)

	cat	car	bat	bar
cat	0	0.0036	6.3972e-004	0.0030
car	-0.0036	0	-0.0042	-6.4958e-004
bat	-6.3972e-004	0.0042	0	-0.0036
bar	-0.0030	6.4958e-004	0.0036	0

3 Results-Conclusions

The calibration method used in this study allows for the examination of distributed, but potentially subtle, differences in representations of semantic information between mechanistic encoding of the language and the EEG.

It was noted that all comparisons in table 2 gave significant differences, which outcome can be considered reliable and promising for further investigations. It should be noted that the words with the same suffix as bat-cat and bar-car showed more consistency. This observation is in agreement with research in the field of neuroscience, which indicates that it is "the syntactically relevant word category information in the suffix, available only after the word stems which carried the semantic information" [23].

These multivariate techniques offer advantages over traditional statistical methodologies in linguistics and neuroscience. The proposed method creates a new basis in the measurements of writing because, for the first time, a code of the digital lexical-word, such as ASCI code, is attempted to calibrate based on a biological signal. The experiments gave helpful results. Consequently, this method will be tested along with the classification step by appropriate neural network classifiers. The proposed metrics have been implemented in the Matlab Language.

In conclusion, the proposed method differs from all existing methods of semantic decoding EEG because it aims to build a model that explains how an acoustic signal lexical content may be shaped so that it can form the basis of linguistic education of the brain. In other words, the proposed model is based on a different logic in relation to aforementioned studies because it creates a combination of two scientific areas, which are neuroscience and linguistics.

References

1. Todirascu, A., Romary, L. & Bekhouche, D: Vulcain—an ontology-based information extraction system. Natural Language Processing and Information Systems pp. 64–75 (2002).
2. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. & Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society for Information Science, vol. 41, pp. 391–407 (1990).
3. Wang, J.Z. & Taylor, W.: Concept forest: A new ontology-assisted text document similarity measurement method. (2007).
4. Rodríguez, M.A. & Egenhofer, M.J.: Determining semantic similarity among entity classes from different ontologies. IEEE transactions on knowledge and data engineering, vol. 15 (2), pp. 442–456. (2003).
5. Corley, C. & Mihalcea, and R.: Measuring the semantic similarity of texts. In: Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, pp. 13–18, (2005).
6. Samsonovic, A.V., Ascoli, G.A. & Krichmar, J.: Principal Semantic Components of Language and the Measurement of Meaning. PloS one, 5 (6), e10921 (2010).

7. Tversky, A. & Gati, I.: Similarity, separability, and the triangle inequality. Psychological Review, 89, pp. 123-154 (1982).
8. Fauconnier G.: Mental Spaces. Cambridge, UK: Cambridge University Press, (1994).
9. Gardenfors, P.: Conceptual spaces: The geometry of thought. MIT Press, Cambridge, MA (2004).
10. Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, : Handbook of Latent Semantic Analysis. Lawrence Erlbaum Associates, Mahwah, NJ (2007).
11. LePendu, P., Dou, D., Frishkoff, G. & Rong, J.: Ontology database: A new method for semantic modeling and an application to brainwave data. Scientific and Statistical Database Management, pp. 313–330 (2008).
12. Chan, A.M., Halgren, E., Marinkovic, K. & Cash, S.S.: Decoding word and category-specific spatiotemporal representations from MEG and EEG. Neuroimage (2010).
13. Indefrey, P. & Levelt, W.J.M.: The spatial and temporal signatures of word production components. Cognition, 92, pp.101–144 (2004).
14. Canolty, R.T. et al.: Spatiotemporal dynamics of word processing in the human brain. Frontiers in neuroscience, 1, p. 185 (2007).
15. Murphy, B. et all. EEG decoding of semantic category reveals distributed representations for single concepts. Brain and Language (2011).
16. Poeppel, D. & Embick, D.: Defining the relation between linguistics and neuroscience. In: Twenty-first century psycholinguistics: Four cornerstones, pp.103–118 (2005).
17. Vapnik, V.N. The nature of statistical learning theory. Springer Verlag (2000).
18. Cristianini, N. & Shawe-Taylor, J.: An introduction to support Vector Machines: and other kernel-based learning methods. Cambridge University Press, (2004).
19. Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N. & Watkins, C.: Text classification using string kernels. The Journal of Machine Learning Research, 2, pp. 419–444 (2002).
20. Box, G.E.P. Jenkins, G.M. & Reinsel G.C.: Time Series Analysis Forecasting and control. Wiley John Wiley & Sons, Inc, (1970).
21. Stone, M. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. Journal of the Royal Statistical Society. Series B (Methodological), 39, pp. 44–47 (1977).
22. Poulos, M., Rangoussi, M., Alexandris, N. & Evangelou, A.: Person identification from the EEG using nonlinear signal classification. Methods of Information in Medicine, 41, pp. 64–75 (2002).
23. Friederici, A.D., Gunter, T.C., Hahne, A. & Mauth, K.: The relative timing of syntactic and semantic processes in sentence comprehension. NeuroReport, 15, pp. 165 (2004).