# Combining fMRI Data and Neural Networks to Quantify Contextual Effects in the Brain

Nora Aguirre-Celis(✉)[1,2] and Risto Miikkulainen[2]

[1] ITESM, E. Garza Sada 2501, Monterrey, NL, 64840, Mexico
[2] The University of Texas at Austin, 2317 Speedway, Austin, TX, 78712 USA
{naguirre,risto}@cs.utexas.edu

**Abstract.** Does word meaning change according to the context? Although this hypothesis has existed for a long time, only recently it has become possible to test it based on neuroimaging. Embodiment theories of knowledge representation suggest that word meaning consist of a collection of attributes defined in terms of various neural systems. This approach represents an unlimited number of objects through weighted attributes and the weights may change in context. This paper aims at quantifying such dynamic meanings using computational modeling. A neural network is trained with backpropagation to map attribute-based representations to fMRI images of subjects reading everyday sentences. Backpropagation is then extended to the features, demonstrating how they change in different sentence contexts for the same word. Indeed, statistically significant changes occurred across similar contexts and across different subjects, quantifying for the first time how attribute weightings for the same word are modified by context. Such dynamic representations of meaning could be used in future natural language processing systems, allowing them to mirror human performance more accurately.

**Keywords:** Context Effect, Concept Representations, fMRI Data Analysis, Neural Networks, Embodied Cognition

## 1 Introduction

Embodiment theories of knowledge representation [1-3] propose that word meaning consist of a set of features, or attributes, that represent the basic elements of meaning. This approach provides an efficient method for representing an unlimited number of object types through weighted attributes. Recently it has become possible to ground this theory to brain imaging, mapping the semantic attributes to different brain systems. In particular, Binder et al. [4] identified a distributed large-scale brain network linked to the storage and retrieval of words. This brain network was used as the foundation for the Concept Attributes Representation (CAR) theory. CAR theory propose that words are represented as a set of properties that are basic components of meaning. Additionally, these properties are grounded in different neural systems such as sensory,

motor, visual, spatial, temporal, affective, and others, based on the way concepts are experienced and acquired [4-7].

An intriguing challenge to such theories is that concepts are dynamic, i.e. word meanings are not fixed entries or lists of attributes, but dynamically processed each time a word is encountered [8]. For example, a pianist would invoke different aspects of the word piano depending on whether he will be playing in a concert or moving the piano. When thinking about a coming performance, the emphasis will be on the piano's function, including sound and fine hand movements. When moving the piano, the emphasis will be on shape, size, weight and other larger limb movements. It is possible to track the dynamic meanings of words by measuring how the attribute weighting changes across contexts.

The research stream of this paper aims to quantify this phenomenon through computational modeling. A neural network is trained to map brain-based semantic representations of words (CARs) into fMRI data of subjects reading everyday sentences. Backpropagation is then repeated separately for each sentence, reducing the remaining error by modifying only the CARs at the input of the network. As a result, the strengths of the attributes in the CARs change according to how important each attribute is for that sentence context.

Previous work with the available fMRI data set resulted in semantically meaningful changes. These changes were reported anecdotally in [9]. Word meaning was represented as a collection of attributes (CARs), grounded in observed brain networks. In two separate experiments, Multiple Linear Regression and a nonlinear Neural Network were used to map the CARs to the FMRI data in order to understand how the CARs could change to approximate the actual sentence representations seen in fMRI images. The results suggested that different features of word meaning were activated in different contexts. The linear mapping approach yielded disorganized results but the nonlinear mapping characterized the results in a meaningful manner.

In this paper, the CARs changes were analyzed more systematically. Interesting context effects were observed for different shades of meaning. Also, the changes in the CAR representations were averaged across subjects, and found to be statistically significant. In fact, the FGREP model captured the context of the sentence combining the meaning of the individual words. Based on this process, in the future it may be possible to create the word meaning dynamically in a natural language processing system, making it more sensitive to the semantic nuances that humans perceive and use.

The CARs theory is first reviewed, and the sentence collection, fMRI data, and word representation data described. The FGREP model is presented, followed by the experiments and how they were tested for statistical significance with the emphasis on aggregating context across subjects.

## 2 Concept Attribute Representation Theory

CARs represent the basic components of meaning defined in terms of observed neural processes and brain systems [4-7]. They are composed of a list of well-known modalities that correspond to specialized sensory, motor and affective brain processes,

systems processing spatial, temporal, and casual information, and areas involved in social cognition. They capture aspects of experience central to the acquisition of event and object concepts (both abstract and concrete). For example, concept ratings on visual and sensory components include brightness, color, size, shape, temperature, weight, pain, etc. These aspects of mental experience model each word as a collection of a 66-dimensional feature vector that captures the strength of association between each neural attribute and the word meaning. Figure 1, shows the CAR for the concept *bicycle*.

The attributes were selected after an extensive body of physiological evidence based on two assumptions: (1) All aspects of mental experience can contribute to concept acquisition and consequently concept composition; (2) Experiential phenomena are grounded on neural processors representing a particular aspect of experience. For a more detailed account of the attribute selection and definition see [4-7]. Section 3.3 describes how the CAR ratings are acquired.

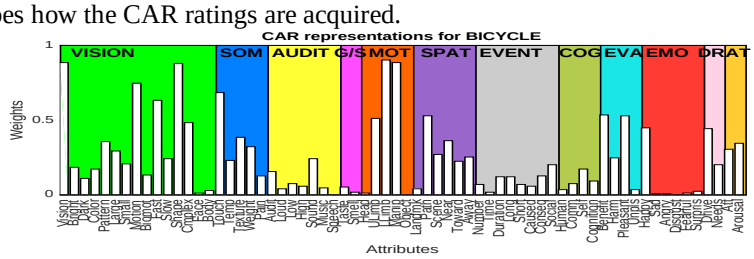



**Fig. 1.** Bar plot for the 66 semantic features in CAR theory. The ratings represent the basic features of *bicycle*. Given that is an object, it gets low weightings on human-related attributes: face, speech, head, and emotion and high weightings on visual, shape, touch, manipulation, and others.

## 3 Data Collection and Processing

Three data sets were used for this study: the sentence collection prepared by Glasgow et al. [10], the Semantic Vectors (CAR ratings) for words obtained via Mechanical Turk [7,11], and the fMRI images of the same sentence collection assembled by the Medical College of Wisconsin [11].

### 3.1 Sentence Collection and Semantic Word Vectors

The sentence set was prepared for use with neural data as part of the IARPA Knowledge Representation in Neural Systems (KRNS) Program [10]. The 240 sentences are composed by 2-5 content words from a set of 242 words (141 nouns, 39 adjectives and 62 verbs). The words were selected toward imaginable and concrete objects, actions, settings, roles, state and emotions, and events, (e.g. *couple*, *author*, *boy*, *theatre*, *hospital*, *desk*, *red*, *flood*, *damaged*, *drank*, *gave*, *happy*, *old*, *summer*, *chicken*, *dog*).

The 242 words (CAR) ratings were collected through Amazon Mechanical Turk [7,11]. In a scale of 0-6, the participants were asked to assign the degree to which a given concept is associated to a specific type of neural component of experience (e.g. "To what degree do you think of a *bicycle* as having a fixed location, as on a map?").

Approximately 30 ratings were collected for each word. After averaging all ratings and removing outliers, the final attributes were transformed to unit length yielding a 66-dimensional feature vector (Figure 1). Note that in this manner, the richness and complexity of representations is based on a direct mapping between the conceptual content of a word and the corresponding neural representations (stimulating perceptual features of the named concept), unlike other systems where the features are extracted from text corpora and the meaning is determined by associations between words and between words and contexts [12-14].

### 3.2 Neural Images

Sentences were presented word-by-word using a rapid serial visual presentation paradigm, with each content word exposed for 400ms followed by a 200ms inter-stimulus interval. Eleven subjects took part in this experiment producing 12 repetitions each. Participants viewed the sentences on a computer screen word by word while in the scanner. The data was acquired by the Center for Imagining Research of the Medical College of Wisconsin [11]. The fMRI voxels were preprocessed and transformed into a single sentence fMRI representation per participant (by averaging all the repetitions), with a final selection of 396 voxels per sentence on a scale from 0.2-0.8, for further use in the computational model.

### 3.3 Data Preparation

Because the neural data set did not include fMRI images for words in isolation, a technique developed by Anderson et al. [11] was adopted to approximate them. The voxel values for a word were obtained by averaging all fMRI images for the sentence where the word occurred. Thus, the vectors include a combination of examples of that word along with other words that appear in the same sentence. The final vector representations became the list of Synthetic Words (called SynthWord) . Because of the limited number of combinations, some of these vectors became identical, and were excluded from the dataset.

Given the final selection of 237 sentences and 236 words (138 nouns, 38 adjectives and 60 verbs), the next step was to identify pairs of contrasting sentences with differences and similarities such as live mouse vs. dead mouse, family celebrated vs. happy family, and playing soccer vs. watching soccer. A collection of 77 such sentences, with different shades of meaning for verbs, nouns and adjectives, as well as different contexts for nouns and adjectives was assembled. This data set was used to prompt Words of Interest during the experimental process (Table 1).

## 4 Computational Model

The technique for analyzing fMRI data is based on the FGREP neural network (Forming Global Representations with Extended BP, [15]). The neural network is trained to predict fMRI sentences (Figure 2), by mapping CARWord (word attribute

ratings) to SynthWord (fMRI synthetic words). SynthWords are combined, to form SyntSent for the predicted sentence, by averaging all words in the sentence. The SynthSent is then compared to the actual fMRISent (original fMRI data), to form a new error signal.

The trained neural network is thus utilized to determine how the CARWords should change in the context of the sentence. That is, for each sentence, the CARWords are propagated and the error is formed as before, but during backpropagation, the network is no longer changed. Instead, the error is used to change the CARWords themselves (which is the FGREP method; [15]). This modification can be carried out until the error goes to zero, or no additional change is possible (because the CAR attributes are already at their max or min limits). Eventually, the revised CARWord represents the word meaning in the current sentence.

For the experiments, the FGREP model was trained 20 times with different random seeds for each of the eleven fMRI subjects. A total of 20 different sets of 786 context-based word representations were thus produced for each subject. In the experiments, the mean of the 20 representations was used for each word. Specific words to be analyzed (i.e. words of interest), were hand-picked from the list of contrasting sentences described in section 3.3, to evaluate the performance of the model and the learned context-based representations. The changes for each attribute were evaluated with a paired *t*-test to determine which ones were statistically significant at the 95% level.

**Table 1.** Sentences examples with differences and similarities in meaning. For instance, the role of the noun *soldier* is used in two different contexts, delivering medicine (good) vs. kicking the door (as an aggressive behavior).

| SEMANTIC CONTRAST | | SENTENCES |
|---|---|---|
| GOOD | 94 | *The soldier delivered the medicine.* |
| AGGRESSIVE | 112 | *The soldier kicked the door.* |
| ANIMAL | 203 | *The yellow bird flew over the field.* |
| | 207 | *The duck flew.* |
| OBJECT | 210 | *The red plane flew through the cloud.* |
| BAD PEOPLE | 119 | *The dangerous criminal stole the television.* |
| | 152 | *The mob was dangerous.* |
| NATURE | 99 | *The flood was dangerous.* |

## 5 Results

The goal of the experiments was to characterize the changes that occur when a word is used in the context of a sentence. Since the importance given to individual attributes of a word varies with context, three analyses to visualize those changes are included in this paper: (1) Characterizing the effect of similar context on two different words, (2) Characterizing the effect of two different contexts on the same word, and (3) Characterizing differences in two contexts. The first experiment analyzed the similarities and differences between the concepts *boat* and *car* across sentences, indicating that they are distinct members of the same category of vehicles. The second experiment examined the conceptual noun-verb combination using the representations of *bird flew* vs. *plane flew*, to evaluate how they result in different degrees of animacy. The third experiment quantified the emotional context of *laughed* and *celebrated* by analyzing how context emerges from thematic associations, [16], and demonstrating

how such cognitive content can be a powerful source of context beyond the more obvious physical context.

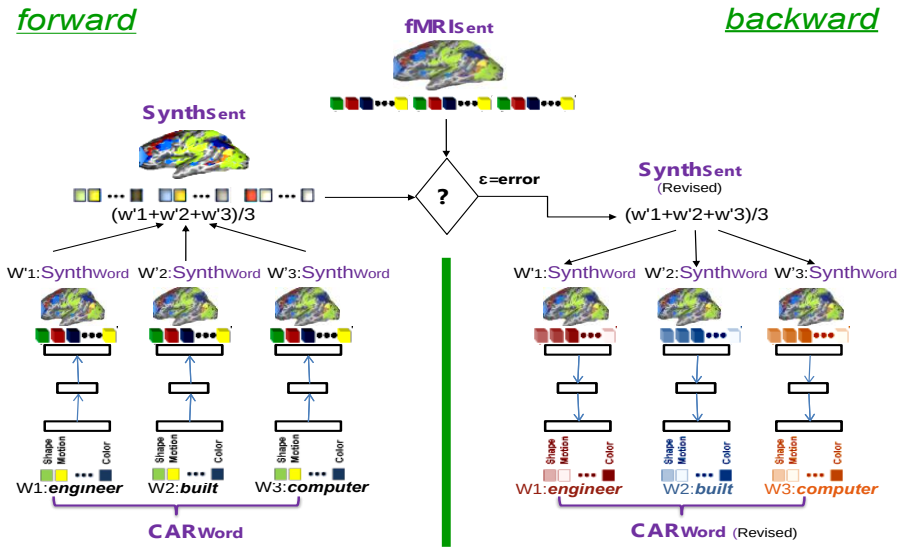



**Fig. 2.** The FGREP model to account for context effects. (1) Propagate CARWord to SynthWord. (2) Construct SynthSent by averaging the words into a prediction of the sentence. (3) Compare SynthSent against Observed fMRISent. (4) Backpropagate the error with FGREP for each sentence, freezing network weights and changing only CARWord. (5) Repeat until error reaches zero or CARs reach their upper or lower limits. The FGREP model captures context effects by mapping brain-based semantic representations to fMRI images.

### 5.1 Effects of Similar Context

In the first experiment the salient attributes for the words *boat* and *car* are compared under the semantic category of transportation vehicles as expressed in 57: *The boat crossed the small lake* and 142: *The green car crossed the bridge*. In principle, *boat* and *car* should be in the same sentence context, but due to data availability, the experiment was designed with sentences that were similar and typical of those nouns. In CAR theory the activation of attribute representations is modulated continuously through attention and the interaction with context. Context draws attention to a subset of attributes, which are then enhanced, forming the basis for object categories. FGREP model quantified such enhanced representations for *boat* and *car*, revealing common underlying properties in the transportation vehicle category [7]. Due to space constrains, only two words are analyzed in this section, but different words were considered (*bicycle* vs. *plane; dog* vs. *mouse; horse* vs. *fish; tea* vs. *water*), with comparable results.

Figure 3 shows the results, averaged across subjects. For *boat* in sentence 57, there are changes on Vision, Large, Motion, Shape, Complexity, Weight, Sound, Manipulation, Path and Scene and event attribute Away, reflecting a large moving

object. Evaluation and Emotion attributes of Benefit, Pleasant and Happy represent the experiential and personal nature of using a boat. Similarly, *car* in sentence 142 shows analogous activation for the same brain areas. Since both belong to the same semantic category, they share the similar context-related attribute enhancement. However, the distinctive weighting on these attributes sets them apart. The FGREP model was thus able to identify the effect of similar context on these two concepts across subjects.

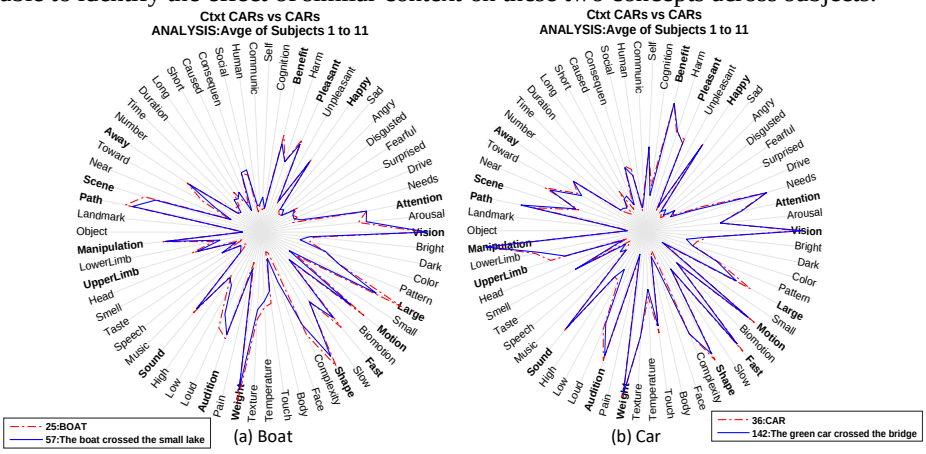



**Fig. 3.** The effect of similar context for the words *boat* and *car* averaged across subjects. Results are shown for the new CARs as an average of all subjects. The dotted lines indicate the original CARs and solid lines specify the context-based representations. Both plots display similar changes but the different weightings set them apart.

### 5.2 Effects of Different Contexts

In this experiment, the attributes of the noun-verb combination are analyzed for the word *flew*, as expressed in 200: *The yellow bird flew over the field*, and 207: *The red plane flew through the cloud*. According to CAR theory, noun-verb interactions arise within multiple brain networks, activating similar brain zones for both concepts. These interactions determine the meaning of the concept combination [7]. Since *bird* is a living thing, animate sensory, motor, affective, and cognitive experiences are activated, including attributes like Face and Speech. In contrast, *plane flew* has salient activations along animate dimensions such as Emotion, Cognition, and, Attention. Figure 4, shows the differences for *flew* in the two contexts. On the left side, all the 66 attributes are displayed and on the right side, only the statistically significant attributes. Due to space constrains, only Subject 9701 is presented, but the analysis was completed on all subjects and for diverse conceptual combinations (*small camera* vs. *small church*; *dangerous criminal* vs. *dangerous flood*; *injured horse* vs. *injured person*; *horse walked* vs. *person walked*), with similar effects.

The results demonstrate context-dependent changes on Sentence 200 with salient activations on animate attributes like Face, Small, and Body, Pain, Audition, Music, Speech, Taste and Smell, as well as Communication and Cognition. In the other hand,

Sentence 207 yields large changes on Vision, Color, Size, and Shape, Weight, Audition, Loud, Sound, Duration, Social, Benefit, Harm, and Attention. These results suggest that FGREP was able to determine the effect of two different contexts into the resulting CARs. As the context varies for each sentence representation, the overlap on neural representations create a mutual enhancement, producing a sharp difference between animate and inanimate contexts.

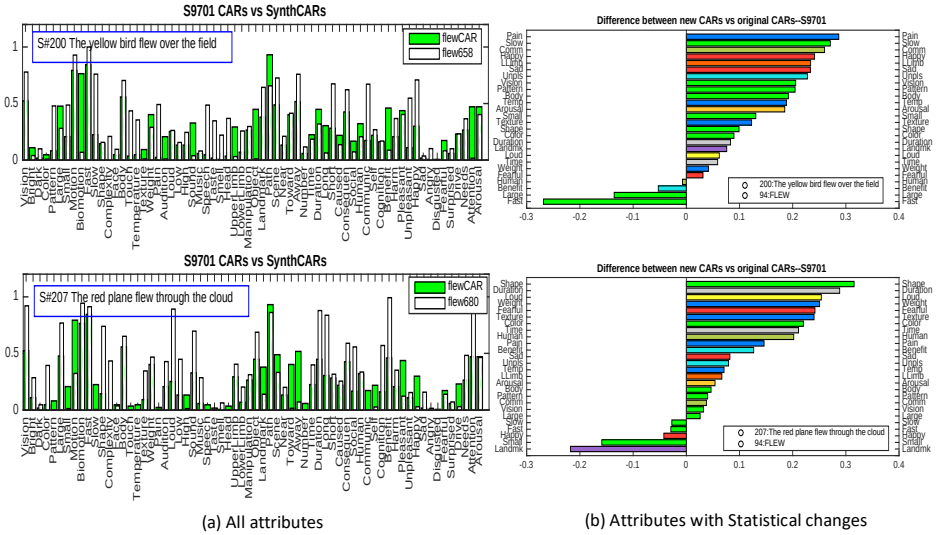



**Fig. 4.** The effect of two different contexts for the word *flew*. (a) Changes of the original CAR vs. new CAR for all 66 attributes. (b) The statistically significant attributes in descending order. The new CARs for Sentence 200 have salient activations on animate features, presumably denoting *bird* properties like Pain, Small, and Communication. Sentence 207, has high activations on inanimate object features, describing a Loud, Large, and Heavy object such as a *plane*.

## 5.3 **Characterizing Differences in Context**

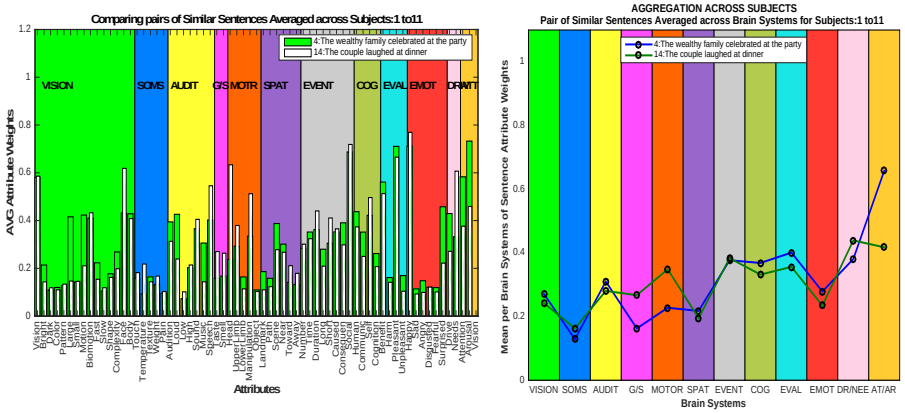
The third experiment examined the common emotional context in Sentences 4: *The wealthy family celebrated at the party*, and 14: *The couple laughed at dinner*, by how such cognitive content can be an instrumental source of context and demonstrating how context develops from external relations.

Many concepts such as *celebrated* and *laughed* refer to affective states and emotions, and other cognitive experiences. One advantage on using CARs is that such experiences count as much as sensory-motor experiences in grounding conceptual representations. When people "feel happy", they experience this phenomenon the same way as the sensory or motor events, except that the perception is internal. Similarly, to evaluate context in these sentences, CAR representations alone cannot capture the thematic associations between concepts (i.e., party, celebration, birthday cake, candles, laugh) unless additional sources provides it. Hence, the third experiment was designed

to quantify that sort of context developed from external relations, i.e., spatial and temporal co-occurrence of events, captured by FGREP.

Figure 5 shows that these sentences resulted on very similar contexts, emphasizing Scenes, Events, and positive Emotions. Figure 5(a) shows the context CARs averaged for each sentence for all subjects. Both sentences are mostly similar on Spatial, Event, and Emotion attributes. Figure 5(b) aggregates these dimensions across the 12 corresponding brain areas according to the CAR theory. All subject brain signatures mainly differ in Gustatory, Motor, and Attention, possibly highlighting that laughing at dinner involves food and requires more head and upper body movements. In contrast, celebrating demands more Attention and Arousal. The results thus suggest that FGREP captures the thematic relations where the two contexts intersect semantically. They also validated that emotional content is a prominent and potentially powerful factor in sentence context, and there are subtle differences in it that can cause subtle differences in word meanings.

Finding how sentence meaning is represented in the brain remains a major challenge [17]. The results in this experiment are significant because they indicate that FGREP captures the thematic knowledge of the sentences by mapping the heteromodal semantic representations (CAR) to fMRI data. By doing so, it is possible to look at the weightings of the brain systems for the entire sentence (as was done in Figure 5b), although the thematic associations exposed by the model require further review.




(a) Celebrated vs. Laughed (b) Averaged

**Fig. 5.** Results featuring differences between two contexts averaged across subjects. (a) A comparison of the averaged attributes for each sentence representing *celebrated* and *laughed*. (b) Aggregation analysis across subjects for each brain zones. These context-based representations differ mostly in the Gustatory, Motor, and Attention zones, possibly emphasizing that laughing at dinner involves food and requires more movement than celebrating at the party, but the latter demands more Attention and Arousal.

## 6 Discussion and Further Work

The experiments in this paper suggest that different aspects of word meaning are weighted differently in distinct contexts, and it is possible to identify those changes for

individual concepts, a combination of concepts, and for sentences by analyzing the corresponding fMRI images through the FGREP model. The changes in the CAR representations were averaged across subjects and found to be statistically significant. This result is remarkable considering that the dataset was not originally designed to answer the question of dynamic meaning. Limited by the data available, the experiments presented here address specific cases, however, by expanding the collection (e.g., identical contexts and contrasting contexts) the number of potential observations would increase, making it possible to test more systematically.

Synthetic words built by combining sentences where the word occurs is similar to many semantic models in Computational Linguistics [13,14,18]. Also, synthetic words formed by fMRI sentence representations has been successful in cases like predicting brain activation [11,17]. Although this study does not have a large set of sentences, the FGREP process of mapping semantic CAR words to the synthetic words and further to sentences fMRI refined the synthetic representations by removing noisy information. Still, fMRI images for individual words instead of having to synthesize them, should amplify the observed effects.

Ongoing research is exploring aggregation analysis across sentence contexts. The goal is to determine how similar sentences cause similar changes in word representations. The process starts by forming clusters of the 237 sentence representations. For each cluster, all new CAR representations with similar roles are identified and the changes between the new and the original CARs averaged and correlated with differences between clusters.

In the future, context dependent representations could be utilized in building artificial natural language processing systems. It may be possible to train e.g. a neural network to predict how meaning changes in context. Such a network could be then used as part of an engineered natural language processing system, dynamically modifying the vector representations for the words to fit the context. Such a system should be more effective and more robust in its inference, and match human behavior better.

## 7 Conclusion

Concepts are dynamic; their meaning depends on context and recent experience. In this paper, word meaning was represented as a collection of attributes (CARs), grounded in observed brain systems. The FGREP Neural Network was trained to map CAR representations of words to fMRI images of subjects reading everyday sentences. Backpropagation was then extended to the CAR features, demonstrating how they change in different sentence contexts for the same word. The changes in the CAR representations were averaged across subjects and found to be statistically significant. In the future it may be possible to create such representations dynamically in a natural language processing system, making it more sensitive to the semantic nuances that humans perceive and use.

**Acknowledgments.** We would like to thank Jeffery Binder (Medical College of Wisconsin), Rajeev Raizada and Andrew Anderson (University of Rochester), Mario Aguilar and Patrick

Connolly (Teledyne Scientific Co.) for providing this data and insight for this research. This work was supported in part by IARPA-FA8650-14-C-7357 and by NIH 1U01DC014922 grants.

## References

1. Regier, T.: The Human Semantic potential. MIT Press, Cambridge, Massachusetts (1996).
2. Landau, B., Smith, L., Jones, S: Object Perception and Object Naming in Early Development. Trends in Cogn Sci vol. 27, pp. 19-24 . (1998).
3. Barsalou, L. W. : Grounded Cognition. Annl Review of Psyc., vol. 59, pp. 617-845 (2008).
4. Binder, J. R., Desai, R. H., Graves, W. W., Conant, L. L.: Where is the semantic system? A critical review of 120 neuroimaging studies. Cereb. Cortex, vol. 19, pp. 2767-2769. (2009).
5. Binder, J. R., Desai, R. H.: The neurobiology of semantic memory. Trends Cognitive Sci, vol. 15(11), pp. 527-536. (2011).
6. Binder, J. R., Conant L. L., Humpries C. J., Fernandino L., Simons S., Aguilar M., Desai R.: Toward a brain-based Comp. Sem. Cog. Neuropsychology, vol. 33:3-4, pp. 130-174. (2016).
7. Binder, J. R.: In defense of abstract conceptual representations. Psychon. B & R, 23. (2016).
8. Pecher, D., Zeelenberg, R., Barsalou, L. W.: Sensorimotors simulations underlie conceptual representations: Modality-specific effects of prior activation. Psychon. B & R, vol. 11, pp. 164-167. (2004).
9. Aguirre-Celis, N., Miikkulainen R.: From Words to Sentences & Back: Characterizing Context-dependent Meaning Rep in the Brain. In Proc.39th Annual Meeting of the Cognitive Science Society, London, UK, pp. 1513-1518. (2017).
10. Glasgow, K., Roos, M., Haufler, A. J., Chevillet, M., A., Wolmetz, M.: Evaluating semantic models with word-sentence relatedness. (2016). arXiv:1603.07253
11. Anderson, A. J., Binder, J. R., Fernandino, L., Humpries C. J., Conant L. L., Aguilar M., Wang X., Doko, S., Raizada, R. D.: Perdicting Neural activity patterns associated with sentences using neurobiologically motivated model of semantic representation. Cerebral Cortex, pp. 1-17. (2016). DOI:10.1093/cercor/bhw240
12. Burgess, C.: From simple associations to the building blocks of language: Modeling meaning with HAL. Behavior Research Methods, Inst. & Com., vol. 30, pp. 188-198. (1998).
13. Landauer, T. K., Dumais, S. T.: A solution to plato's problem: The latent semantic analysis theory. Psychological Review, vol. 104, pp. 211-240. (1997).
14. Vinyals, O., Toshev, A., Bengio, S., Erham, D.: Show and Tell: A New Image Caption Generator. (2015). arXiv:1506.03134v2
15. Miikkulainen, R., Dyer, M., G.: Natural Language Processing with Modular PDP Networks and Distributed Lexicon. Cognitive Science, vol. 15, pp. 343-399. (1991).
16. Estes Z, Golonka S, Jones L. L. Thematic thinking: The apprehension and consequences of thematic relations. Psychology of Learn and Motiv, vol. 54, pp. 249-294. (2011).
17. Anderson, A. J., Lalor, E. C., Lin, F., Binder, J. R., Fernandino, L., Humpries C. J., Conant L., Raizada, R. D., Grimm, S., Wang, X.: Multiple Regions of a Cortical Network Commonly Encode the Meaning of Words in Multiple Grammatical Positions of Read Sentences. Cerebral Cortex, pp. 1-16. (2018). DOI:10.1093/cercor/bhy110
18. Mitchell J, Lapata M.: Composition in distributional models of semantics. Cogn Sci. Vol.38, Issue 8, pp. 1388–1439. (2010). DOI: 10.1111/j.1551-6709.2010.01106.x