Lexical Semantic Analysis Based on Cassirer's Philosophy of Symbolic Forms

Lijuan Wei[1, 2]

[1]*Postdoctoral Programme of Philosophy, Dalian University of Technology, Dalian, 116024, China*
[2]*College of Foreign Languages, Handan University, Handan, 056005, China*

Li Li

Applied Foreign Languages Department, Shijiazhuang Vocational College of Scientific and Technical Engineering, Shijiazhuang, 050800, China

Abstract
As a kind of vocabularies learned after the human language acquisition process, verb is characterized by certain complexity, whose meaning interpretation requiring nouns, adverbs and other underlying word involved. Its semantic meaning mainly refers to an action event, which can be contained in dynamic videos. However, semantic analysis in being is unable of mastering the dynamic performance of verbs. The paper defines the semantic expression structure of Cassirer's Philosophy of Symbolic Forms, in which frame and argument are included. The frame is used to organize the cognitive structure of contextual knowledge, and the argument is controlled by the frame to achieve description of specified context. As per this verb semantic definition, the verb semantic acquisition model ViMac-V based on Cassirer's Philosophy of Symbolic Forms is established. The experiment proves that ViMac-V is efficient in extracting elements of the frame and argument. There are totally 5 groups of frames and 4 argument parts of speech (62 argument vocabularies) on 1 word are acquired.
Key words: SEMANTIC MEANING; ARGUMENT, CASSIRER'S PHILOSOPHY OF SYMBOLIC FORMS, VIDEO INFORMATION, DYNAMIC PERFORMANCE, CONTEXT KNOWLEDGE

1. Introduction

In the 1980s, because that Cassirer's An Essay on Man translated and published in our country, the cultural philosophy of Cassirer used to be a hottest topic during China's fever of cultural philosophy studies. The most important cultural philosophy works written by Cassirer is Philosophy of Symbolic Forms and his earlier works related to the epistemology were not translated into Chinese. As a result, China's depth studies on Cassirer's cultural philosophy thoughts were affected; especially those affecting the transition from recent philosophy to modern philosophy have not received enough attention and researched. The subsequent translation versions have many problems as well [9, 10]. The paper analyzes Cassirer's philosophy of symbolic forms, and both re-analysis and researches on training corpus selection, corpus processing and pretreatment, feature selection, semantic

meanings modeling and training are required.

In language acquisition tasks based on video information [1, 2, 3], previous studies mainly have delved into video information relation between nouns and adjectives [4]. However, visual semantic acquisition of verbs has seldom been concerned, primarily caused by the complexity of verb concept and its corresponding visual scenes [5]. For example, the concept that the contemplation of philosophy is not referring to thoughts, but the independent thought of an individual. Satish and Et al. [6] acquired the image schema representation relating to two transitive verbs "close" and "away" by the relative motion events of two moving objects [7], and used the image schema representation to describe state of running motion in another real scene. However, these all focus on connection process of verb and visual feature, and losing sight of an appropriate representation to verb semantic meaning. In this way, selected verb objects are all transitive verbs with simple argument structures, such as verbs "push" "throw" "close" etc. A descriptive statement formed by these verbs only has a basic form of "SVO" and verb only acts as a conjunction between subject and object. In this case, the acquired verb semantic meaning actually has no essential distinction with nouns and adjectives in mode of visual representation.

For this purpose, the paper proceeds with verbs from linguistics structure and visual information. The semantic expression structure of verbs based on linguistics frame is defined, in which frame and argument are included. Secondly, the connection between verb arguments and visual information of formal training corpus is formed by using a self-organizing neural network and a videotext description of certain scale. Finally, a semantic representation reflecting the verb language structure and relying on visual information is established.

2. Data preparation

The video-based Meaning Acquisition of Chinese Verb ViMac-V is a supervised learning model structure. ViMac-V adopts videotext description as the training sample, conducts corresponding pretreatment respectively to videos and texts by requirements of the verb semantic definition based on frame semantics [8], connects video and text information with the help of self-organizing mapping network, and establish the verb semantic representation based on visual features. ViMac-V training data preparation is conducted following visual and linguistic information, mainly including the collection and production of video files and description files. In this way, it may provide visual scenes and description language used for training and testing to subsequent SOM net.

2.1 Description of Video Corpora

The video corpora can be divided into two parts: real scene corpora and virtual scene corpora. The real scene corpora mean using vision devices to capture motion scene in real physical environment, in which the description of moving objects' visual features should be obtained by calculation. Real corpora are used in testing process to evaluate the descriptive power to real motion scene. According to states in linear displacement, the production of virtual corpora selects 5 parameter scopes to determine the scene of a certain linear displacement. Table 1 presents 5 parameter scopes and related values.

Table 1. Controls Parameter List of Video Corpora Production

Parameter Scope	Variable Parameter	Parameter Choice	Parameter Specification
Direction	lr, rl, du, ud	values corresponded 1, 2, 3, 4	representing four moving directions: up, down, left, right
Speed	a, b, c	divided into three levels with values corresponded 12, 24, 48, representing 0.5 sec, 1 sec and 1.5 sec for moving 0.1 on the screen	representing the ball moving speed
Start Position	x, y	horizontal and vertical ordinate values the ball locates on the screen, the screen area is 1*1	start position the ball moves on the screen
Destination Position	x, y	horizontal and vertical ordinate values the ball locates on the screen, the screen area is 1*1	destination position the ball moves on the screen
Moving Distance	dist	increased by unit step with value of 0.2	distance the ball moves on the screen

2.2 Artificial Corpus Annotation

The annotated corpus is similar to a teaching language heard when human seeing the visual scene in the language acquisition process, referring to language description of ball displacement event here. The annotated corpus may be acquired in forms of voice of text. Limited by the noise effect in voice-word transition process, it is similar to the corpus preparation with visual features. We use text input method with high accuracy rate in acquisition process and use voice output in the testing process. An adult with normal cognitive competence is chosen as an annotator. The produced video files may be divided into 5 groups, 200 videos are contained for each, and 1000 are annotated as training corpora. Wherein self-developed Corpus Building is adopted as annotation tool. No limitation may exist in the annotated statements in principle. All annotated statements are natural cognitive reactions annotators have after watching videos, and the created description statements are in the form of open natural language.

3. Processing of text-described alignment corpora

The alignment corpus described by videos/texts is designated to acquire dual-channel information comprised of visual and language features and provides to the subsequent self-organizing neural network group. The dual-channel information formed by modal characteristics is similar to the teaching scene and teaching language perceived in the human language acquisition. For instance, human gradually acquire visual meanings of related languages by utilizing the continuous co-occurrence of visual and language features with cognitive significance.

3.1 Video Feature Analysis

The video feature analysis is aimed at spatial and temporal features related to motion concept from video corpora. The visual feature of description motion object selected by this paper is comprised by a 12-dimensional feature vector, i.e.:

$$V_f = \left[sn, x_i, y_i, x_f, y_f, s, d_x, d_y, sp_x, ep_x, ep_y \right]^T$$

The physical significance of dimensional visual vector in V_f is relatively simple. The Sudoku coordinate system is to divide object motion area equably into 9 blocks of 3X3 and use the block label to replace coordinate values of the motion object. The Sudoku coordinate system has data reduction effect on start and destination coordinate, preventing the case that later orientation concept may be affected due to the excessive even of start and destination coordinate on the object motion area.

3.2 Underlying Words Selection Based on Visual Features and Lexical Co-Occurrence

To classify word units derived by pretreatment to correspond 4 cognitive components of the displacement frame, i.e. 4 argument word categories. When it comes to division of 4 argument word categories, underlying words in these 4 argument word categories need to be determined firstly as the classification foundation. To compare underlying words with other words to obtain the eventual classification result. Two standards may be relied on selecting underlying words: 1) count of word frequency 2) visual features and co-occurrence rate of words. The calculation of the latter one shall use parameter values of video files production as visual features. If visual feature set, is the word set of annotated corpora W, then

$$V = \{v_1, v_2, \cdots, v_n\}, W = \{w_1, w_2, \cdots, w_n\}.$$

Wherein the occurrence frequency of the visual features $v_i (1 \le i \le n)$ in the file is a_i, the occurrence frequency of words $w_j (1 \le j \le m)$ is w_j , w_j, the occurrence frequency in the same column is c_{ij}, then the co-occurrence of v_i and x_j shall be calculated as below:

$$p_{ij} = \frac{c_{ij}}{\left(a_i - c_{ij} + 1\right)\left(b_j - c_{ij} + 1\right)}$$

The number was less in speed, so word frequency is not taken into consideration. At the same time, some unrelated words are artificially deleted in the above results.

3.3 Calculation of Word Similarity Based on Mixed Measurement

The calculation for similarity of words in the corpora and various underlying words is based on the mixed measurement of minimum throughout distance and part-of-speech distance. The distance measurement equation of word w_i and underlying word w_j is presented as below:

$$D\left(w_1, w_2\right) = \alpha \cdot \left(D_{MED}\left(w_1, w_2\right) + 1\right)^{-1} + \beta \cdot D_{POS}\left(w_1, w_2\right)$$

3.4 Frame Extraction Based on Bigram Model

By the measurement and division of argument part-of-speech, 4 types of words are obtained, i.e. $W_A^h \left(h \in [speed, dir, start, des]\right)$. Different argument words have different statement characters. They reflect highlight perspectives in visual cognition of action events through the activation in frame, such as cognitive attention attracted by speed significance and direction significance. These perspectives can be achieved by selecting and sorting the argument part-of-speech in frame. The frame in corpus is extracted by bigram statistical model, and the transition probability of two

argument part-of-speeches is counted up by using the forward algorithm.

$$P\left(W_A^i \middle| W_A^j\right) = < W_A^i, W_A^j > / W_A^j$$

The probability the argument part-of-speech acting as beginning or destination of sentence can be calculated by the following equations:

$$P\left(W_A^i \middle| begenning\right) = < W_A^i, begenning > / n \quad P\left(W_A^i \middle| ending\right) = < ending, W_A^i > / n$$

4. Experimental results of argument part-of-speech and frame extraction

This part is experimental results of argument part-of-speech and frame extraction for annotated corpora. The argument part-of-speech and frame obtained shall be acted as f dual-channel language and vision information, and shall be be deemed as input signal for subsequent frame activation and argument categorization.

4.1 Classification Results of Argument Part-Of-Speech

The classification results and performance of argument part-of-speech can be seen in Table 2. Calculated according to the results in Table 2, the total F value of these three argument part-of-speeches is 86.9%, satisfying essentially the requirement to be input signal for subsequent frame activation and argument categorization.

Table 2. Classification Results of Argument Part-Of-Speech

Argument Part-Of-Speech	Total	Number of Wrongly Classified	Number of Missed Classified	Correct Number	Accuracy Rate	Recall Rate
Direction Argument	512	2	213	510	99.6%	70.5%
Orientation Argument	931	264	2	667	71.6%	99.7%
Speed Argument	610	0	63	610	99.8%	90.6%

According to the analysis of results in Table 2, the main source of wrong classification is as below. 1) Annotators made spelling and sentence mistakes. 2) The mistakes of segmentation and introduction may exist. 3) Because the basis of algorithm used by part-of-speech annotation is the context, one word with several part-of-speeches may occur and may thus lead to inaccurate similarity calculation. 4) Annotated words and phrases excluded in direction, orientation and speed argument but classified in above argument part-of-speech may exist. Through artificial modification to above mistakes, we eventually obtained the extracted argument part-of-speech list, as shown in Table 3.

Table 3.

Argument Part-Of-Speech	Argument Words
Direction Argument	Left to right, to right, right to left, up, right-hand to up, down, up, left to right, right
Orientation Argument	End, middle, left, right side, to end, from middle, left and right, right-hand, left-hand, left and right, lower right, upper left, up, left to right, upper left, upper right, top, center, left side, bottom, lower right, under, lower left, tdestination bottom, horizontal, middle, center, left, upper middle, up, upper, top, vertical, down, bottom, down, upper middle, lower middle, down, right, lower right, bottom, top, lower left, lower middle
Speed Argument	Fast, very slow, slower, slow, slowly, very fast, rapid

In these three types of argument part-of-speeches, motion orientation part-of-speech (dass2) shall be further divided into start position W_A^{start} and destination position. It is because grammar position and order of the two are different in annotated statements and frame extraction. By adopting regulation, the subdivision defines preposition words before “to” is start position words and after is destination position words. Finally, four groups of argument part-of-speech are obtained: $W_A^{dir}(class1)$、W_A^{start}.

$(class2), W_A^{speed}(class3)$, $W_A^{des}(class4)$

The subsequent frame extraction is then carried out.

4.2 Results of Frame Extraction

After completing argument part-of-speech classification, the frame extraction shall be calculated by the Bigram. The state transition diagram after the calculation is shown as below.

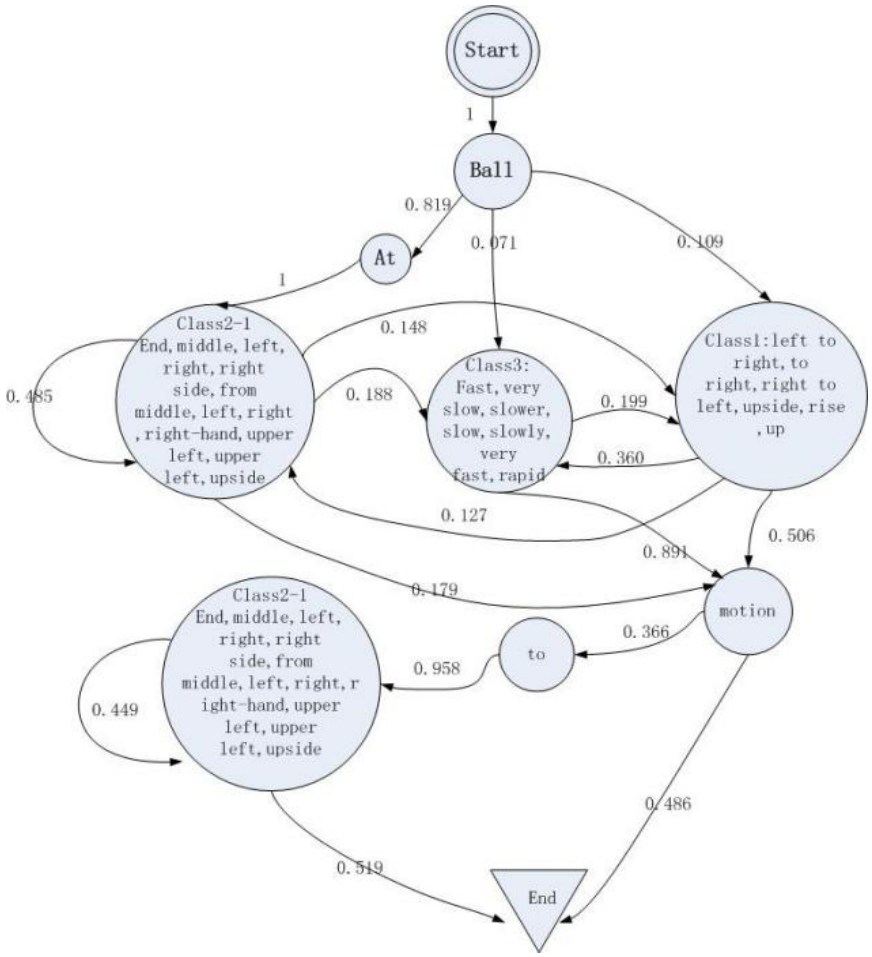

Figure 1. Bigram Transition Probability Diagram

We consider the grammar structure with maximum probability output in Figure 1 as the universal verb frame, i.e. each argument part-of-speech may use it, or all visual cognitive perspectives may be activated. However, for corresponding frames of the specific verbs, because each cognitive perspective activated is different, for example, the frame of verb "fall" may have something to do with the start and destination points of the direction and orientation argument. The difference between of "fly" and "skin" may reflect the occurrence of speed argument in the frame. As a result, we extract the frame of each verb by using the above Bigram model. The result is shown in Table 4.

Table 4. Results of Verb Frame Extraction

Verb Frame	Sample Size	Argument grammar order	Frame type	Corresponding Verb
F_0	1000	$W_A^{start} W_A^{speed} W_A^{dir} W_A^{des}$	Universal Frame	Any Verb
F_1	404	$W_A^{start} W_A^{dir} W_A^{des}$	Type of Direction	Exercise
F_2	915	$W_A^{start} W_A^{dir}$	Type of Direction	Move, Fly
F_3	140	$W_A^{start} W_A^{speed}$	Type of Speed	Fly, Skim
F_4	387	$W_A^{start} W_A^{speed} W_A^{des}$	Type of Speed	Fall, Fly to

It is regulated that the universal frame F_0 may serve for any verbs when outputting the statement, which is replaced by "exercise" or "move" in most cases. 7 remaining verbs serve for 4 frames. These 4 frames can further divided into two groups as per different center arguments. The argument located in the middle of start position may be deemed to represent the cognitive significance of the frame. As a result, because the activated argument W_A^{dir} is divided into the direction frame, then F_2, F_2 serve for "move" "exercise" "fly". F_3、F_4 is divided into the speed frame, which is not use W_A^{dir}, but start argument W_A^{start}, W_A^{des} to express the motion direction, this word shall be divided into speed frame. So far, we have completed the pretreatment of the annotated corpora and obtained the required motion visual features and components of frames and arguments related to verb semantic structure. Wherein the visual feature contained with 12-dimensional visual vector r to constitute (equation (4-1)). 4 types of argument component words and 5 verb frames are obtained from annotated statements, which are $\{W_A^{speed}, W_A^{dir}, W_A^{start}, W_A^{des}\}$, $\{F_0, \cdots, F_4\}$, respectively. The superscripts of four types of argument component words correspond to speed, dir, start and des. 5 frames represent 5 different verb grammar structures. Visual and language information forms different dual-channel input by feature selection for corresponding frame activation mechanism and implementation of argument component categorization for subsequent self-organizing neural network.

In the experiment of argument word categorization, selecting the related combination with visual features and language features of each sub-network in SOM etwork group as the input signal of sub-network categorization shall be done in the first place. Table 5 is feature combination selected from dual-channel information of 4 sub-networks.

Colored hexagons are inserted between neuron nodes in SOM network by U chart to represent the distances between neuron nodes trained.

Table 5. Feature Combination Selected from Four SOM Sub-Networks

	Aligned Corpora Number	Selected Visual Feature	Argument Component Word
Speed Sub-Network	612	$[x_i, y_i, x_f, s]$	Adverbs describing motion speed, such as "fast" "slow"
Direction Sub-Network	663	$[x_i, y_i, x_f, d_x, d_y]$	prepositional phrases or multi-word expressions describing directions, such as "left to right" "up"
Start Sub-Network	881	$[x_i, y_i, ns_x, ns_y]$	prepositional phrases, nouns or multi-word expressions describing directions, such as "up" "from middle" "top of the left hand"
Destination Sub-Network	445	$[x_i, y_i, ne_x, ne_y]$	prepositional phrases, nouns or multi-word expressions describing destination position, such as "end" "top"

Colored hexagons are inserted between neuron nodes in SOM network by U chart to represent the distances between neuron nodes trained. The bar chart along refers to the relation between color and distance. Deeper color represents closer distance between the trained neuron weight vectors in the feature space. On the contrary, shallower color represents farther distance between trained neuron weight vectors in the feature space. As a result, in the U chart, shallow colored tape always represents classification boundary. According to speed component sub-net works in Figure 2(a), it can be seen clearly that they are divided into three categories, however, classification information of the three remaining sub-networks are not significant. In order to obtain the specific hierarchical category information on the U chart, the clustering processing requires to be conducted to obtain the corresponding clustering results. Wherein 3 categories are obtained by speed components, 4 categories are obtained by direction components, 9 categories are obtained by start and destination components.

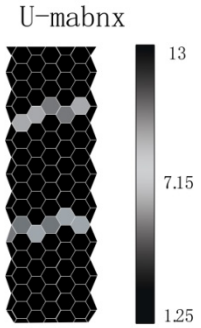

Figure 2(a). U Chart of speed sub-Network

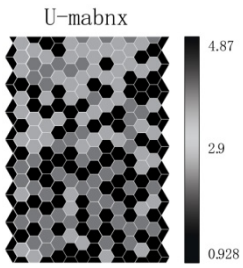

Figure 2(b). U Chart of Direction sub-Network

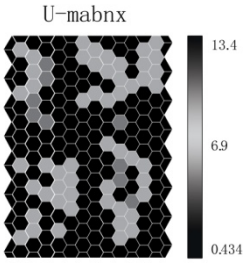

Figure 2(c). U Chart of Start sub-Network

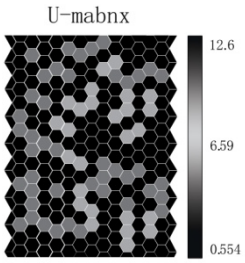

Figure 2(d). U Chart of Direction sub-Network

In order to test the influence SOM-based argument component categorization results exert on the output performance of ViMac-V, we conducted evaluation experiments of description language output. The influence on the performance exerted by component categorization is the sole factor concerned for the time being. As a result, statements are formed by selecting displacement frame with universality of verb frame, i.e. four argument components related to the concept of displacement are all activated, and the difference on displacement modes on verb frame are not considered, verbs used in the generated statements are consistently replaced with the verb "move". In addition, in organizing the whole statement by frames, argument category words obtained require to be added with subject words such as "ball" (the motion subject concept is acquired in static part-of-speech description task and can be directly used). Some auxiliaries are added to organize the video description statements eventually. The video corpora used by evaluation are50 un-annotated AVI videos and description statements generated shall be marked from speed, direction, start, destination, and whole statement by the annotators. The evaluation scores are divided into three levels, 1 score for correct, 0 score for mistake and 0.5 for uncertain. Table 5-3 has shown 5 description statement examples to the generated testing videos. According to the different selection mode of start and destination argument features, the description statements obtained from the Sudoku labels and natural words. Table 6 has given the comparison of eventual testing results. It can be seen that the accuracy rate of speed argument is approximately 75%. The result difference between the two columns is caused by twice evaluations of the same group of speed argument words. Direction argument outputs are correct indeed. As for output words of start and destination argument, the results adopted by natural words and Sudoku labels have a greater difference. Relating start argument words, the accuracy rate of statements adopted by natural words is 65%, while the accuracy rate of statements adopted by Sudoku labels is 90%, an increase of 25%. Relating the destination argument words, the accuracy rate of statements adopted by natural words is 56.67%, while the accuracy rate of statements adopted by Sudoku labels is 88.33%, an increase of 31.66%. Relating whole statements, the accuracy rate of statements adopted by natural words is 51.67%, while the accuracy rate of statements adopted by Sudoku labels is 85.33%, an increase of 33.66%. As a result, through the evaluation, basically, the description on 50 evaluation videos by ViMac-V accords to the human's cognitive results to videos. The description statements obtained by the Sudoku visual features and language label improving orientation categorization experiment are better than description statements generated by original natural words and the accuracy rate of start, destination, and whole statement are improved about 30%. In the end, small evaluation experiment is conducted to certify that argument categorization experiment is correct and the feature improvement of orientation argument components is valid and essential.

5. Conclusion

As a result, this paper defines the verb semantic representation planning of "frame + argument". In this way, the verb semantics can be together represented by frame and argument category. According to the definition of such representation, ViMac-V needs to extract the related verb frame and argument words from annotated corpora. When it comes to the extraction of argument words, the underlying words shall be selected by visual features and word co-occurrence, and the argument part-of-speech shall be classified by approach of word measurement between parts of speech and minimum edit distance. After obtaining each group of argument part-of-speeches, the verb frame is extracted by bigram model. Eventually, 4 groups of argument part-of-speeches, 5 groups of frames, and 7 verbs are obtained. The connection between verb semantics and video information is achieved through frame activation mechanism and argument word categorization based on SOM network.

Table 6. Corresponding Relationship between Natural Language Words and Sudoku Labels of Start Argument

Sudoku Labels of Start Argument	Natural Language Words of Start Argument
up #9	middle#2;top of upper middle #1;top #2
down #6	upper right in the middle #1;bottom right in the middle #1;upper left in the middle #1;bottom left in the middle #1;lower right in the middle #1;lower left in the middle #1
left #7	down #1;upper left #1;lower right #1;top #1
right #6	Right #1;end #1;lower right #2;bottom right #1
middle #7	upper middle #1;lower right in the middle #1;lower left in the middle #1;upper left in the middle #1;upper right in the middle #1;
upper left #7	left #2;upper left #1;upper left #1
lower left #6	down #1;lower middle #2;bottom left#1
upper right #7	upper middle #1
lower right #5	end #2;lower right #2;bottom right #1;upper middle in the right #2

Acknowledgments

This article belongs to the subject of Hebei Institute for education research, subject name: University English Major Linguistics Course Teaching Mode of Hebei Province, subject number: 13040986, anchorperson: Wei Lijuan.

References

1. Siskind J.M. (2014) Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *Journal of Artificial Intelligence Research*, 59(20), p.p. 31-90..
2. Roy D. (2013) Grounding words in perception and action; computational insights. *Trends in Cognitive Sciences*, 29(8), p.p. 389-396.
3. Pastra K. (2009) Viewing Vision-Language Integration as a Double-Grounding Case. *American Association for Artificial Intelligence*, 60(6), p.p.58-87.
4. Yu, C. and D.H. (2014) On the integration of grounding language and learning objects, *Proceedings of the Nineteenth National Conference on Artificial Intelligence*, Paris, France, p.p. 320-340.
5. Gentner H. (2009) Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. *Language Development*, 20(12), p.p.650-680.
6. Mukeijee, G. (2013) Acquiring Linguistic Argument Structure from Multimodal Input using Attentive Focus, *IEEE International Conference of Development and Learning*, Monterey, U.S.A, p.p. 201-230.
7. Johnson M.(2005) *The Body in the Mind: The Bodily Basis of Meaning*, Imagination, and Reason, University of Chicago.
8. Fillmore C.J. (2012) Frame semantics, in Linguistics in the Morning Calm, *Journal of Linguistic Society of Korea*, 46(3), p.p. 111-137.
9. Kohonen T. (1981) Automatic formation of topological maps of patterns in a self-organizing system, *Conference on Image Analysis*. Helsinki, Finland, p.p.120-125.
10. Ernst C. (1968) *The Philosophy of Symbolic Forms*, Yale University Press.