Modeling the Phenomenological Concepts for Figurative Processing of Natural-Language Constructions

Oleg Bisikalo[0000-0002-7607-1943], Yuriy Ivanov[0000-0003-2125-1004],
Vladyslava Sholota[0000-0002-7073-8727]

Vinnytsia National Technical University Vinnytsia, Ukraine

obisikalo@gmail.com, y.ivanov@gmail.com,
vladislava.sholota@gmail.com

Abstract . An approach to the formalization of the «meaning pyramid» notion for natural-language constructions (NLC) on the basis of the power set graphical interpretation is proposed in the paper. Made on the power set basis, formal interpretation of the associative pairs cognitive space is considered as a distributive bonded lattice with a complement to each element. The two-base algebraic structure is denoted as the Boolean algebra of the sense (BAS). With the BAS the following set-theoretic determinations for the phenomenological concepts of the NLC figurative processing are defined: syntagma, humor anchor, text, attention focus, such types of memory as associative memory, random access memory, memory cache.

Keywords : power set; natural-language constructions; sense; Boolean algebra; phenomenological concepts; syntagma; humor anchor; memory.

1 Introduction

The complexity of the problem of the text information semantic processing, which arises under the influence of the need of taking into account many phenomenological factors [1], can be somewhat reduced at the expense of choosing the most simple modeling methods [2]. Modeling such notions as semantic space [3] and associative pairs [4] seems to be the most promising in this way. For the first time, in the paper [5] there was justified the application of a well-known from the set theory concept of a power set which was used for the formal interpretation of the cognitive space of linguistic images (LI) associative pairs. Thus, a special class of graphs is introduced, in which there is taken into consideration the specifics of the semantic combination of word forms as LI verbal signs on the basis of statistical analysis of their relationships in the natural-language constructions (NLC).

Modeling of the phenomenological components of NLC and their interaction is important in view of solving the actual problem of recognition of humor, which can be considered as AI-complete. Detection in some context of a certain degree of humor is a complex problem in the processing of natural language [6]. Solutions of this problem are connected with the simulation of the concept of so-called humor anchor, which leads to causing a humorous effect [7] as a very important concept of semantic analysis of the text.

Thus, a succeeding development of the formalisms proposed in [5] for modeling important concepts of semantic text analysis is relevant. Herewith, we will understand, according to [8], that language image is a set of all the words relating to the chosen lemma of some natural language. In this case, natural language constructs are an undirected graph has combined LI, for example, in a frame of the single sentence.

If suggested approach to compare with the syntactic analysis for in-stance, which also provides syntactic groups and relations among them - in the NLC definition, we aggregate many syntactic groups with similarity meaning. In our opinion, this way will help to go from traditional sentences to sentences with figurative meaning, as the term LI suggests a priory.

Therefore, we would like to find the answer in this study to the main question - whether it is possible to model in the framework of set theory the key phenomenological concepts for the figurative processing of NLC according to research results [8].

The *aim of the research* is to denote the formal set-theoretic determinations of key phenomenological concepts for the figurative processing of NLC: syntagma, humor anchor, text, attention focus, and such types of memory as associative memory, random access memory, cache as well.

2 Introduction of the «meaning pyramid» formal concept

Let's consider the possibility of interpreting meaningful concepts of semantic analysis of textual information by ignoring the direction of the associative relation between the LI pair and its statistical weight. We will denote the power set of a set I as P (I). According to the definition [9] a P (I) lattice is a shortened designation for

$$P(I) = \{Y \mid Y \subseteq I\}. \qquad (1)$$

If any finite set X has n number of elements, it goes without saying that $P(I)$ has 2^n elements (and from here goes the name of the concept "power set"). It is proposed to depict the power set on the plane in a graph form, which illustrates the relation of the partial order of the lattice and consists of $n+l$ layers.

For the $P(I)$ there will be used the relation of an order with a following definition: if $<I, \leq>$ is a finite partially ordered set, $I \leq Z$ is equivalent to a $I = I_1 \leq I_2 \leq ... \leq I_n = Z$ chain, where each I_{i+1} cover I_i. Any finite ordered set I can be represented as a visual scheme, with all its elements given in the form of graph vertices (or nodes). In this case a Z vertices will be located above an I vertices only if $I \leq Z$, and if Z covers I , I and Z are connected by a line segment.

Such schemes are used [10] not only for depicting certain partially ordered sets, but also for creating primary definition of partially ordered sets. In the second mentioned case, an order relation, according to its definition, is a relation that connects elements-vertices of the scheme with polygonal chains.

Power set $P(I)$ of the consisting of n number of elements image set I is convenient to represent as a set of binary sequences (numbers), built according to the following rules:

- the number of digit positions of each number equals to n;
- the number of all numbers is 2^n;
- if an *i-th* element of the I is incorporated into the $P(I)$ subset, then there is 1 or 0 in the *i-th* digit position of the corresponding code;
- an empty set $\varnothing$ is also a part of the $P(I)$, and it is represented as $000\ldots0$ (the number of digit positions is n).

We will consider the graphic image of the $P(I)$ power set, which meets all the stated above requirements, to be a unit cube of an n-dimensional space (in the n-dimensional coordinate system). The regularities of the growth of the n-dimensional space, which can be detected on the basis of this approach, allow us to develop a recursive algorithm for the $P(I)$ power set graphic interpretation. In fact, the problem is reduced to the calculation of the coordinate projection of each point of the n-dimensional space on a two-dimensional plane in the case when $n > 2$.

Using the method of mathematical induction and the analysis of cases with small values of n, we are to define the following relations for a multi-layered graph projection of a power set lattice (n-dimensional polyhedron) on the plane:

- the number of layers: (from 0 to n);
- the number of vertices: 2^n;
- the number of relations from the i-th layer vertex:
- up: $n-i$;
- down: i;
- the number of the i-th layer vertices (it equals to the numbers of combinations from n to i) [11]:

$$C_n^i = \frac{n!}{i!(n-i)!}; \qquad (2)$$

- the total number of the connections between the vertices of the graph:

$$\sum_{i=0}^{n} C_n^i \cdot (n-i) = \sum_{i=0}^{n} \frac{n!(n-i)}{i!(n-i)!}. \qquad (3)$$

Figure 1 illustrates a graphic interpretation of the power set for the 4-dimensional lattice.

According to the logic of the NLC figurative analysis model [8], there is the only vertex on the zero (ground) graph layer (it can be represented as a code $000\ldots000$ or as a symbol Θ) which will be considered as a global zero. There are always n vertices with 100…000, 010…000, 001…000, …, 000…010, 000…001 codes on the first layer (these codes are possible language images which form a cognitive space of associative pairs).

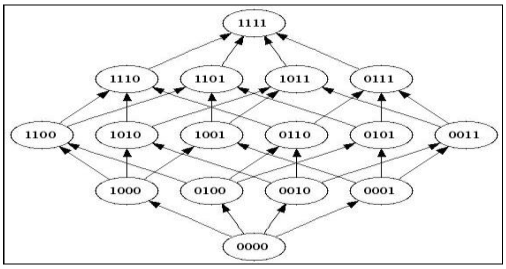

Fig. 1. Multi-layer graphic projection of the power set lattice on the plane for the case when *n* =4.

There is also always a set of all possible associative pairs on the second layer; these associative pairs can be created from *n* LI (designation of each vertex has ones in two digit positions and zeros in all other digit positions of the code). By the same scheme on the *i-th* layer, the number of units in the code of each vertex is equal to *i*, and, according to the logic of the NLC figurative analysis model, vertices from the third one and up to the n-th layer, correspond to the sentence and / or whole texts. It is clear that there is one vertex on the n-th layer with ones in all *n* digit positions of the code. It is to be considered as a global one and be designated as Δ .

Made on the power set basis, formal interpretation of the associative pairs cognitive space we will consider as a distributive bonded lattice with a complement to each element [11].

The main problem of the method of natural language information encoding, which is proposed on the basis of the power set, is in a computational complexity. It applies not only to the vertices, especially in the middle layers of the power set, but also, what is more important, to the links (connections) of the given graph. Figure 2 shows a fragment of a Boolean graph for the case n = 10, obtained using the *GraphViz* package in the automatic mode of arrangement of vertices on plane.

That is why the additional information about the domain knowledge is needed, because it will contribute to the reduction of the computational complexity through filtering out unacceptable search options.

As it is shown in the work [5], an accumulation of the associative memory connections, which are based on the primary information of events and / or sentences from natural language texts, is an extremely important characteristic of the text information figurative processing model. Let’s introduce a formal sentence model as a syntagma (according to [12]). In order to solve the problem, it is proposed to introduce additional data about the parameters of the pairs subset union into a syntagma, the time of syntagma recording, the type and the direction of the connections in each pair, the type and rules of representation of each language image according to its role in a certain associative connection. Moreover, the model must determine and maintain the value of the weight of the direct and reciprocal associative connection in each pair, what was first proposed in [8].

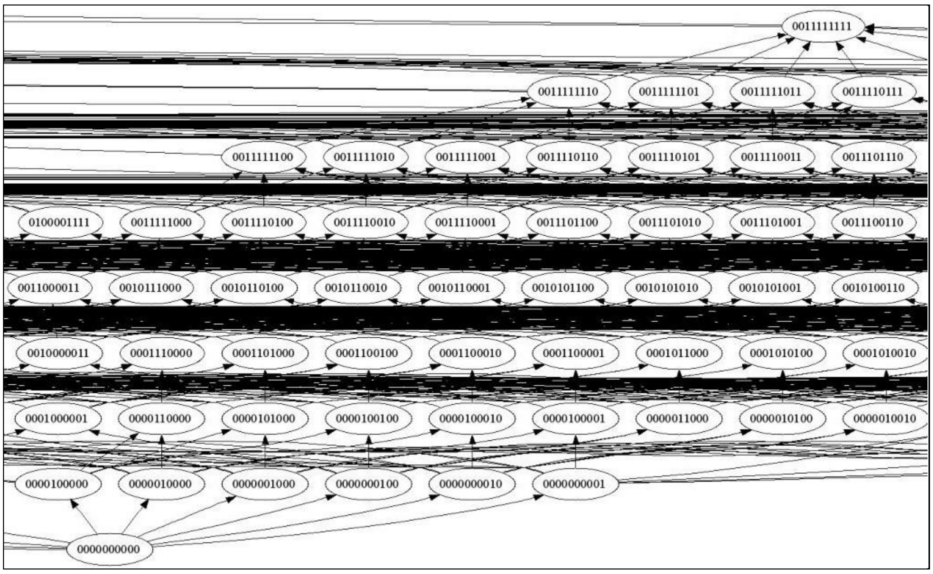

Fig. 2. Fragment of the Boolean graph for n = 10

A formal ability of building a so called “meaning pyramid” for a certain syntagma on the basis of the information represented on the second layer will be introduced. The concept of the “meaning pyramid” was firstly introduced in the paper [8] in contrast to the similar concept of growing pyramidal networks [13] when considering the construction of a linguistic processor on the basis of the NLC figurative sense model. The marking of the vertices and links between them (these links belong to a separate “meaning pyramid” and are marked on the figures in bold italics and lines) on the power set graph is made in the accordance with the following algorithm:

1. To activate all the vertices which correspond to the LI of a certain syntagma from the first layer.
2. To activate on the second layer only those vertices which correspond to the associative pairs of the given syntagma, i. e. only the composition and the direction of the connections between language features of the images of a certain event are fixed in the pyramid.
3. On all the layers (higher than the second one) of the power set there should be activated the vertices, the codes of which are the results of the union operation of the codes of any pair of vertices activated on the previous layer. But only in one from all the digit positions of the taken code one can substitute zero.
4. For the whole “meaning pyramid” it is necessary to activate only those connections which go from the activated vertices of the i-1st layer to the activated vertex of the i-th layer.

With the help of the “meaning pyramid” and by means of processing the LI (syntagma) structures, the sequence of events can be physically inserted into the long-term memory of the system in the form of a LI set (the 1st layer) and in the form of the set of associative links between them (the 2nd layer). But there appears an absolutely new opportunity to renovate formally all the meaningful vertices from the third layer and from all following layers which can be interpreted as the LI combinations which did make sense in the previous experience of the NLC system processing.

3 Boolean algebra of the sense and the formalization of the main concepts of the NLC image processing model

The model of the NLC image processing will be introduced with the help of a two-base algebraic structure [14]. In order to make that, there will be applied the interpretation of the space of associative pairs in the form of a power set graph and with the concept of the "meaning pyramid" taken into account. The two-base algebraic structure will be further considered as the Boolean algebra of the sense (BAS):

$$BAS = \langle B; \Omega_{\mathrm{b}} \rangle, \tag{4}$$

where

$$B = \{Word, Number\} \text{ – bases,} \tag{5}$$

and

$$\Omega_{\mathrm{b}} = \{OP, RE, IF\} \text{ – system signature.} \tag{6}$$

Within the first of the bases of “Word” it is necessary to distinguish the following types of symbolic sequences (words):

$$Word = \{\operatorname{Im} age - Word, Link - Type, \operatorname{Re} - Word\}, \tag{7}$$

The words which represent the language images will be classified according to their roles in a syntagma [8]: $Object - Quality$; $Object$; $Notion$; $Method$; $Method - Quality$

$$\operatorname{Im} age - Word = \begin{Bmatrix} Object - Quality, Object, Notion, \\ Method, Method - Quality \end{Bmatrix}. \tag{8}$$

The words “Link-Type” denote the role of the LI in a sentence that corresponds to a syntagma because of a set of syntagmatic relationships types,

$$Link - Type = \{r1, r2, r3, r4, r5, r6, r7\}, \tag{9}$$

where $r1$ is a definition, $r2$ – predicate, $r3$ – subject, $r4$ – adverbial modifier of place, $r5$ – adverbial modifier of time, $r6$ – adverbial modifier, $r7$ – object. Further verbal details of the main types of syntagmatic relationships can be made in the form of subsets of the corresponding interrogative pronouns (Pronoun).

The words from the set Re-Word will denote the attributes of the relationships which have to be introduced into the system memory block.

As the other base (*Number)*, the numbers of the following types will be considered:

$$Number = \left\{ Bi, Logic, Integer, Time \right\} , \tag{10}$$

where *Bi* is a n-dimensional binary code of the power set lattices; *Logic* – logic value 0 (False) та 1(True):

$$Logic = \left\{ 0,1 \right\} ; \tag{11}$$

Integer – integer non-negative numbers:

$$Integer = \left\{ x \mid x \in Z^{+} \right\} ; \tag{12}$$

Time – real non-negative numbers:

$$Time = \left\{ x \mid x \in R^{+} \right\} , \tag{13}$$

For the n-dimensional binary code, the code with *i* ones will be marked through the *Bi*$_i$ code, and, respectively, *n–i* zeros in all other digit positions. It goes without saying that all the vertices with the *Bi*$_i$ code belong to the i-th layer of the power set graph.

$$Bi = \left\{ Bi_i ,\ i = \overline{1, n} \right\} . \tag{14}$$

Consequently, the two-base algebraic structure BAS allows to use the interpretation of the space of ordered pairs of images in the form of the power set graph, and to consider all the significant concepts of the figurative thinking model as the bases.

The proposed formalism provides with an opportunity to obtain theoretical multi-plane determinations of the components of the figurative thinking conceptual model, despite their complex phenomenological character. Let's put such a formal hierarchy of basic concepts into an algebraic model of figurative thinking [14]:

Syntagma is a set of binary codes Syntagma which corresponds to those vertices of the “meaning pyramid” graph which serves as a figurative representation of events and a formal analogue of a simple declarative sentence (humor anchors can be considered a subset of syntagma sets)

$$Syntagma = \left\{ x \mid x \in Bi_i ,\ \ i = \overline{3, n} \right\} ; \tag{15}$$

Text is a set of binary codes *Text* which corresponds to the power set vertices and represents the combination of some sequence of *m* syntagmas in such a connected subset, where each syntagma has at least one common image with other

$$Text = \left\{ x \mid (x \in Bi_i ,\ \ i = \overline{3, n}), (Bi - Sy_j \cap x \neq \Theta) \right\} , \tag{16}$$

where $Bi-Sy_j \in Syntagma$ – *j*-th syntagma from those united in a text;

Associative memory is a set of binary codes $Assoc-memory$ of all the vertices from the second layer of the power set

$$Assoc-memory=\{x \mid x \in Bi_2\}\ ; \tag{17}$$

Memory cache is a set of binary codes $Super-memory$ which consists of *n* vertices from the first layer of the power set, each of which can be activated at any time as a result of modeling the process of perception of the corresponding image by sensory organs

$$Super-memory=\{x \mid x \in Bi_1\}; \tag{18}$$

Attention focus points at one of the power set vertices with a binary code *Focus* from the set of images $Super-memory$ in each of the discrete time cycles

$$Focus := x \in Super-memory\,; \tag{19}$$

Random Access Memory is a set of binary codes $Oper-memory$ which consists of the vertices from the 5th to the 9th layer of the power set, each of which represents a group of active images at the given point of time

$$Oper-memory=\left\{x \mid (x \in Bi_i,\ \ i=\overline{5,9})\right\}. \tag{20}$$

The task of research has been realized by terms (4)-(20).

Let’s take as an example the sentence «Many people [from] our land move [off]» which consists of 5 generalized elements. The Figure 3 illustrate the power set graph for the example, vertices for subject and predicative of the sentence on the first and the second layers are marked with grey color.

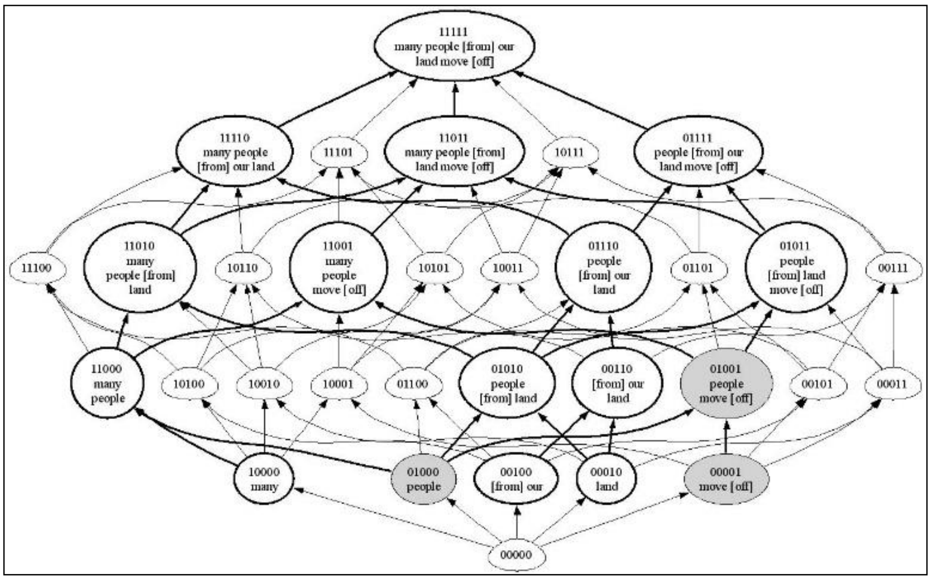

Fig. 3. An example of the power set graph of the syntagma for the case when $n = 5$

The vertice of the graph with the code 11111 can be considered as Syntagma (15), and 4 activated pairs of the second layer of the pyramid, in particular "many people", "people [from] land", "[from] our land" and "people move [off]" make up $Accos-memory$ (17). The peaks of the first layer of the pyramid correspond to $Super-memory$ (18), with the focus of attention (19) of them most likely to fall on one of the gray-marked vertices "people" or "move".

4 Conclusion

Within for the first time proposed algebraic system BAS there were denoted such formal set-theoretic determinations for the following phenomenological concepts of the NLC figurative processing of the model: syntagma, humor anchor, text, attention focus, and such types of memory as associative memory, random access memory, and memory cache as well.

The effectiveness of the proposed formalisms – using of the algorithmic concept of the “meaning pyramid” allows combining the advantages of the NLC elements encoding on the power set basis with the reduction of the computational complexity in processing a needed graph from the NP to the P.

There is one more possibility to evaluation of the suggested approach. If we compare our model with the known method of processing n-grams, then, apart from taking into account all n-grams of the text, we can significantly reduce the dimension of the required database due to:

- the way of combining into one n-gram of a whole set of variations in which words and even word forms are in the different positions;
- the ability to analyze n-grams with a larger value of n as the result of a combination of n-grams with a smaller value of n.

There are no numerical estimates of such a reduction in dimension of the database right now, but this is the subject of further research.

The advantage of the research results compared to the well-known Word Net system is that we can specify the composition of some Synset automatically, without the involvement of linguistic experts. On the other hand, we can to customize a machine-learning model for generating Synsets based on examples from Word Net.

The practical value of the proposed approach – we can be speedy and visually used such an accumulation of the LI pairs in the oriented power set graph that did make sense in the previous experience of the NLC system processing.

So, the scientific hypothesis that was accepted as the result of the conducted study, there is a possibility to model the key phenomenological concepts for the figurative processing of NLC in the framework of set theory.

According to the authors, the further development of the subject is the modeling of humorous phenomena, in particular the processes of recognition and using humorous anchors on the basis of the proposed algebraic approach to the formalization of the concepts of associative figurative thinking of human.

5 References

1. Lund, K., Burgess, C., Audet, C.: Dissociating semantic and associative word relationships using high-dimensional semantic space. In: 18th annual conference of the Cognitive Science Society, pp. 603-608. La Jolla, California (1996).
2. Sahlgren, M.: The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-dimensional Vector Spaces. PhD Dissertation. Stockholm University, Sweden (2006).
3. de las Penas Cabrera, I.: A note on the envelopes of an associative pair. Comm. Algebra **32,** 4133–4140 (2004). doi: 10.1081/AGB-200034014
4. Brown, R., Higgins, Ph., Sivera, R., Nonabelian: Algebraic Topology: Filtered Spaces, Crossed Complexes, Cubical Homotopy Groupoids. EMS Tracts in Mathematics **15** (2011). doi:10.4171/083.
5. Bisikalo, O., Ivanov, Y., Karevina, N.: Encoding of Natural Language Information on the Basis of the Power Set. In: Proceedings of the 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), pp. 17-20. Lviv, Ukraine (2018). doi:10.1109/STC-CSIT.2018.8526732.
6. Attardo, S.: Linguistic theories of humor. Mouton de Gruyter (1994).
7. Yang, D., Lavie, A., Dyer, C., Hovy, E.H.: Humor Recognition and Anchor Extraction. In: Conference on Empirical Methods in Natural Language Processing. Lisbon, Porugal (2015). doi:10.18653/v1/D15-1284.
8. Bisikalo, O., Cięszczyk, C., Yussupova, G.: Solving problems on base of concepts formalization of language image and figurative meaning of the natural-language constructs. In: Proc. SPIE **9816**, 98161U. Lublin, Poland (2015). doi:10.1117/12.2229046.

9. Herzog, J., Hibi, T.: Distributive Lattices Bipartite Graphs and Alexander Duality. In: J Algebr Comb 22: 289 (2005), https://doi.org/10.1007/s10801-005-4528-1.
10. Zaki, M.J.: Scalable algorithms for association mining. In: IEEE Transactions on Knowledge and Data Engineering **12**, issue 3, p. 38 (2000).
11. Griggs, J., Killian, Ch.E., Savage, C.D.: Venn Diagrams and Symmetric Chain Decompositions in the Boolean Latticeм. In: The electronic journal of combinatorics **11**, issue 1, # R2 (2004).
12. Bohland, J., Minai, A.: Efficient associative memory using small-world architecture. In: Neurocomputing **38–40**, pp. 489-496 (2001). https://doi.org/10.1016/S0925-2312(01)00378-2.
13. Gladun, V., Vashchenko, N.: Analitical Processes in Pyramidal Network. In: Information Theories and Application. FOI-COMMERCE, Sofia (2001). http://www.aduis.com.ua/English/public.HTM#V.Gladun,%20N.Vashchenko.
14. Bisikalo O., Lisovenko, A., Jahumovuch, O., Trachenko, S., Pradivliannyi, M.: System of Computational Linguistic on Base of the Figurative Text Comprehension. In: Proceedings of the 2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP), pp. 69-74. House of Lviv Polytechnic National University, Ukraine (2016).