Языкознание
Linguistics

УДК 81–116 http://doi.org/10.32603/2412-8562-2019-5-4-102-114
Оригинальная статья / Original Article

Linguistic Data Model for Natural Languages and Artificial Intelligence. Part 1. Categorization

Oleg M. Polyakov✉
Saint-Petersburg State University of Aerospace Instrumentation, St Petersburg, Russia
✉road.dust.spb@gmail.com

Introduction. The article opens a series of publications on the linguistics of relations (hereinafter R–linguistics), the purpose of which is to formalize the processes studied by linguistics, to expand the possibilities of their use in artificial intelligence systems. At the heart of R-linguistics is the hypothesis that mental and linguistic activity is based on the use of consciousness model of the world, which is a system of specially processed relationships observed in the world or received by consciousness in the process of communication.
Methodology and sources. This article is devoted to the axiomatization of the categorization process. The research methods consist of the development of necessary mathematical concepts for linguistics.
Results and discussion. Axioms of categorization are defined and their equivalence with other systems of axioms is established. The concept of linguistic spaces, which consist of categories formed on the basis of axioms, is formulated. The properties of linguistic spaces are defined. In the paper are introduced the concepts of forming species which are important in decompositions of spaces, and in the transition to a parametric representation and language. Three variants of categorization are considered, the most important of which is verbal categorization. The evaluation of the results and their further development in different directions is carried out.
Conclusion. At the end of the article some additional comments are made for further publications of the series.

Key words: R–linguistics, categorization, linguistic spaces, generators, types.

For citation: Polyakov O. M. Linguistic Data Model for Natural Languages and Artificial Intelligence. Part 1. Categorization. DISCOURSE. 2019, vol. 5, no. 4, pp. 102–114. DOI: 10.32603/2412-8562-2019-5-4-102-114

Conflict of interest. The author declares no conflict of interest.

Received 03.06. 2019; adopted after review 01.07.2019; published online 25.10.2019

Лингвистическая модель данных для естественных языков и искусственного интеллекта. Часть 1. Категоризация

О. М. Поляков✉
Санкт-Петербургский государственный университет аэрокосмического приборостроения
Санкт-Петербург, Россия
✉road.dust.spb@gmail.com

Введение. Статья открывает серию публикаций по лингвистике отношений (далее R–лингвистике) целью которой является формализация процессов, изучаемых лингвистикой, для расширения возможностей их использования в системах искусственного интеллекта. В основе R-лингвистики лежит гипотеза о том, что мыслительная и языковая деятельность базируется на использовании сознанием модели мира, которая представляет собой систему специальным образом переработанных отношений, наблюдаемых в окружающем мире или полученных сознанием в процессе коммуникации.
Методология и источники. Методы исследования заключаются в развитии необходимых математических представлений для лингвистики.
Результаты и обсуждения. Определены аксиомы категоризации и установлена их эквивалентность с другими системами аксиом. Сформулировано понятие лингвистических пространств, которые состоят из категорий, образованных на основе аксиом. Определены свойства лингвистических пространств. Введены понятия образующих и видов, играющих важное значение в разложениях пространств, а также в переходе к признаковому представлению и языку. Рассмотрены три варианта категоризации, важнейшим из которых является глагольная категоризация. Произведена оценка полученных результатов и их дальнейшее развитие в различных направлениях.
Заключение. Сформулированы некоторые дополнительные замечания для связи с дальнейшими публикациями серии.

Ключевые слова: R–лингвистика, категоризация, лингвистические пространства, образующие, виды.

Для цитирования: Поляков О. М. Лингвистическая модель данных для естественных языков и искусственного интеллекта. Часть 1. Категоризация // ДИСКУРС. 2019. Т. 5, № 4. С. 102–114. DOI: 10.32603/2412-8562-2019-5-4-102-114

Конфликт интересов. Автор заявляет об отсутствии конфликта интересов.

Поступила 03.06.2019; принята после рецензирования 01.07.2019; опубликована онлайн 25.10.2019

Introduction. This article opens a series of publications on the linguistics of relations (R-linguistics) – a formal direction in linguistics. It is hardly necessary to say how important any formalization in linguistics. Formalization allows you to have clear definitions and understand the boundaries of what is allowed. In addition, formalization is the basis of any computer application of linguistics. This series of articles aims to build the axiomatic foundations of the model of the world, formed by consciousness and underlying linguistic human activity. The most important first step on this path is the axiomatics of categories, which this article is devoted to. Questions of categorization are presented in detail in cognitive linguistics. Axiomatization allows to reveal fundamental problems of cognitive linguistics, although the article is not enough for such conclusions.

The philosophical foundations, the principles of R-linguistics are the subject of a separate discussion and are beyond the scope of this article. R-linguistics assumes that the main function of our consciousness is the prediction of events, actions, phenomena, etc. This function of consciousness ensures the survival of the species along with other functions: fertility, protection, size, poisons, etc. Every day we make thousands of various predictions, without which our life would be impossible. We predict muscle contractions that allow you to reach the desired touches, predict the next color of the traffic light when approaching the intersection, the behavior of your dog to distinguish it from the actual behavior and determine that the dog is sick, etc. The further and more accurate we predict, the greater our ability to survive.

Prediction is always the product of the work of a certain model and, therefore, the main function of consciousness is the modeling of the surrounding world. R-linguistics uses the most general approach to modeling the world, based on the observation of all kinds of relationships in which the objects around us interact. 0-ary relations are just specific objects of the surrounding world. Unary relations are categories that for various reasons combine 0-ary objects. Binary and in the general case n-ary relations describe the interactions of different numbers of objects. It is necessary to distinguish between categorization, as the process of forming categories, and identification, as the process of relating specific objects to already formed categories. This article deals only with categorization. The issues of categories and signs are not touched here either, as they are related to identification and are considered separately. Categorization only by signs is fruitless: it describes the placement of objects in boxes with different names, and nothing more. Another thing is when signs identify categories obtained for other reasons. In this case, observing the features of objects, one can make predictions of those processes that lie in the basis of one or another categorization. In this sense, the main type of categorization is the verb, because within such a categorization, the identification of objects allows one to make predictions of their behavior (interaction). Of course, not all verbs synthesize categories, but this is the topic for another discussion.

Methodology and sources. As a method of investigation in this paper uses the results of the following mathematical topics: the theory of relations, the lattice theory, general topology, theory of data dependencies. The research materials are presented in the form of mathematical proofs of various properties of the linguistic model associated with the process of categorization.

Results and discussion.

1. Categorization axioms.

Let U be the universe of objects that are categorized. Let θ denotes the operation, which for some set of examples forms a category. This is exactly the operation that was in the head of Robinson Crusoe, who was watching the index finger of Friday. Thanks to it, using a lot of received examples, he guessed which category was being discussed. And at first one should make a few comments.

When we talk about the size of a category, we mean the whole existing set of examples of this category. But, as a rule, we do not know the limits of the category, even if it consists of a finite set of objects. For example, no one knows the set of all the tables of humanity, which, moreover, changes every minute. Therefore, in fact, the operator θ for each specific object determines whether this object is the same category example along with the set of examples given. For example, after demonstrating some examples of trees, Robinson could define the category “tree” and after pointing to trees that were not mentioned by Friday at the beginning. This means that he mastered the operator θ. Under the result of applying the operation θ, we will understand all the many examples that could potentially be obtained if we step by step applied the operator θ to various objects of the universe. Although the operator and the operation θ are different things, we will not take into account this distinction in the future, considering these explanations to be sufficient.

Consider what properties can be in the operation θ. So, let X be a set of objects from U. Obviously, the category θ(X), formed on the basis of the set of objects X, should include these examples X. Of course, after forming the category, other examples will also be included in it, but the initial examples must be included surely. So, we denote this property as an axiom:

$$A1 \qquad X \subseteq \theta(X).$$

The axiom A1 in mathematics is called the axiom of extensiveness. According to the axiom A1, each of the examples on the basis of which a category is formed is necessarily included in this category for the simple reason that is an example of this category. In particular, for the universe U, we have $U \subseteq \theta(U)$. Since, by definition, the universe U contains all the elements, then $U = \theta(U)$. In other words, the universe itself is a category. Let C be any category obtained with the help of the operation θ, and X be some set of examples from C, then

A2 from $X \subseteq C$ follows $\theta(X) \subseteq C$.

Axiom A2 says that any choice of many examples in any category does not allow to go beyond this category. Axiom A2 will be called the axiom of correctness. From this point on, categorization refers to the use of the operation θ with the specified properties [1]. Any process to be called a categorization must at least satisfy the axioms A1 and A2. Somewhat later, we will consider categorization options, and now we will define its properties.

From axiom A2 follows.

Proposition 1. If $X \subseteq \theta(Y)$ and $Y \subseteq \theta(X)$, then $\theta(X) = \theta(Y)$.

The proof directly follows from A2.

In mathematics, the closure operation with the following axioms is well known:

A1 $X \subseteq \theta(X)$ (extensiveness);

C1 $\theta(X) = \theta(\theta(X))$ (idempotency);

C2 from $X \subseteq Y$ follows $\theta(X) \subseteq \theta(Y)$ (monotonicity).

Theorem 2. The systems of axioms <A1, A2> and <A1, C1, C2> are equivalent.

Proof. For the idempotency axiom C1 of A1, it follows that $\theta(X) \subseteq \theta(\theta(X))$. Since $\theta(X) \subseteq \theta(X)$, it follows from A2 that $\theta(\theta(X)) \subseteq \theta(X)$, and, therefore, $\theta(X) = \theta(\theta(X))$. For axiom C2, let $X \subseteq Y$. From the axiom A1 $Y \subseteq \theta(Y)$, and, therefore, $X \subseteq \theta(Y)$. Applying A2, we obtain $\theta(X) \subseteq \theta(Y)$. Conversely, we show that the system of axioms for the closure operator implies A2. Let $X \subseteq \theta(Y)$. Then from C2 it follows that $\theta(X) \subseteq (\theta(\theta(Y))$, and taking C1 into account, we get $\theta(X) \subseteq \theta(Y)$.

The axiom system <A1, A2> (or <A1, C1, C2>) defines the properties of a certain map θ that associates with each subset of examples from U an element of a certain family of categories over U. Consider the properties of this family of categories.

In [2] the following axioms are introduced by Moore for purely mathematical purposes:

B1 U belongs to the family of sets;

B2 intersection of family members is a member of the family.

It should be noted that the axiom system <B1, B2> simply defines a certain family of subsets of U with certain properties, and there is no question of any mapping. Therefore, later in the text we have to talk about equivalency, and not about the equivalence of the axiom system <B1, B2> with the systems <A1, A2> and <A1, C1, C2>.

Lemma 3. For any operation θ satisfying the axiom systems <A1, A2> or <A1, C1, C2>, in the family of its closed sets, the axioms B1 and B2 are fulfilled.

Proof. From A1 it follows that $\theta(X) \cap \theta(Y) \subseteq \theta(\theta(X) \cap \theta(Y))$. On the other hand, $\theta(X) \cap \theta(Y) \subseteq \theta(X)$ and $\theta(X) \cap \theta(Y) \subseteq \theta(Y)$, and, therefore, in accordance with A2 $\theta(\theta(X) \cap \theta(Y)) \subseteq \theta(X)$ and $\theta(\theta(X) \cap \theta(Y)) \subseteq \theta(Y)$. Thus, $\theta(\theta(X) \cap \theta(Y)) \subseteq \theta(X) \cap \theta(Y)$ and, therefore, $\theta(\theta(X) \cap \theta(Y)) = \theta(X) \cap \theta(Y)$. Axiom B1 is a direct consequence of A1, taking into account that the set U is maximal.

Remark 4. The proof of Lemma 3 easily extends to an infinite family of intersections of closed sets in axiom B2. We limited ourselves to only two operands solely because we are only interested in finite category systems.

So, the operation θ satisfying the axiom systems <A1, A2> (or <A1, C1, C2>) defines a category system that satisfies the axiom system <B1, B2>. Conversely, for the axiom system <B1, B2>, one can construct an operation θ satisfying the axiom systems <A1, A2> and <A1, C1, C2>.

Theorem 5. For a given family of sets satisfying the axioms B1 and B2, there exists a unique operation θ satisfying the axiom systems <A1, A2> (or <A1, C1, C2>).

Proof. For an arbitrary X, we define θ(X) as the intersection of all closed sets to which X belongs. For U, from the axiom B1, we get θ(U) = U. From B2 it follows that this definition is correct. By constructing the axiom A1 is being performed. Let $X \subseteq C$. It means that C belongs to the family of intersecting closed sets and, therefore, $\theta(X) \subseteq C$.

Suppose that for the same family of closed sets there exists an operation θ_1, different from θ, but satisfying the system of axioms <A1, A2> (or <A1, C1, C2>). This means that there exists a set X for which $\theta(X) \neq \theta_1(X)$. Since θ and θ_1 form the same family of sets, θ(X) corresponds to one set of the family, and $\theta_1(X)$ to the other one. Since A1 is satisfied for θ_1, $X \subseteq \theta_1(X)$ and, therefore, the closed set θ_1 (X) belongs to the intersection family for θ(X), whence $\theta(X) \subseteq \theta_1(X)$. In addition, from A1 $X \subseteq \theta(X)$ and from A2 for θ_1, we obtain $\theta_1(X) \subseteq \theta_1(\theta(X))$. But θ(X) is a closed set, and therefore, according to C1, $\theta_1(X) \subseteq \theta(X)$. So, $\theta_1(X) = \theta(X)$, which contradicts the assumption.

The investigation. Taking into account Remark 4, the category family for the categorization operation is a complete lattice.

The proof follows from the fact that there is a unity in this family (the category of U). In addition, for any subset of categories, there is an exact lower bound (the intersection of categories) [3]. Thus, in addition to the intersection operation in the family of categories, the addition operation is defined: if S and T are two categories, then $S + T = \theta(S \cup T)$. Since the lattice is complete, this operation can also be extended to an infinite number of operands.

Remark 6. Lemma 3 and Theorem 5 make clear the meaning that is embedded in the equivalency of the axiom system <B1, B2> with the systems <A1, A2> and <A1, C1, C2>. They uniquely associate a category family with a categorization operation: one uniquely defines the other. From here we get a simple way to define a category system: you must first define some family of initial categories, including U, and then add to it all possible missing intersections of the family elements. This will uniquely determine the categorization operation that satisfies the axiom system <A1, A2> (or <A1, C1, C2>).

Hence the obvious, but important.

Theorem 7. Let θ_1 and θ_2 be two categorization operations on U. There is also a unique categorization operation θ_3 for θ_1 and θ_2 on U such that the categories of operations θ_1 and θ_2 are categories of operation θ_3, and every other operation with this property forms more categories than the operation θ_3.

Proof. Combine the categories of operations θ_1 and θ_2. If, according to Remark 6, to these categories add all missing intersections, then we get the family of categories of the operation θ_3. This operation will be the smallest, since any operation that combines the categories of original families must, according to B2, contain intersections of the categories of these families.

Remark 8. Operation θ_3 forms new categories in the most economical way, without creating "extra" categories, therefore everywhere in the future, the mixing of categorization operations means a categorization operation obtained by the method described above [4].

Similarly, if θ_1 and θ_2 are two categorization operations on U, then you can enter θ_3 – the **multiplication** of these operations. The set of categories of operation θ_3 consists of all categories that are simultaneously members of families of categories by operations θ_1 and θ_2. Obviously, this is the correct definition of a multiplication operation. Indeed, the universe U belongs to the families of categories for θ_1 and for θ_2, which means that it will be included in the set of categories for θ_3. Similarly, if K and L are two categories of operation θ_3, then by definition they are categories for operations θ_1 and θ_2, and therefore the intersection of these categories will belong to θ_3.

Remark 9. And again, for the operations mixing and multiplication, an infinite number of operands can be considered.

Remark 10. The topological spaces on which modern spatial representations are based are also defined by a special type of closure operation. It is given by the following axioms [5]:

$$A1 \quad X \subseteq \theta(X);$$
$$D1 \quad \theta(X \cup Y) = \theta(X) \cup \theta(Y);$$
$$D2 \quad \theta(\varnothing) = \varnothing;$$
$$C1 \quad \theta(X) = \theta(\theta(X)).$$

From the axioms indicated, for example, this property of the topological closure operation follows:

$$C2 \quad \text{from } X \subseteq Y \text{ follows } \theta(X) \subseteq \theta(Y).$$

Axioms A1, C1, C2 are the axioms of the closure already considered by us, so the topological spaces are a special case of families of categories that satisfy the axioms <A1, A2> (or <A1, C1, C2> or <B1, B2>). By analogy with topological spaces, for convenience, we will call families of closed sets satisfying the axiom systems <A1, A2> (or <A1, C1, C2> or <B1, B2>), *linguistic spaces* (or simply *spaces*). Just like linguistic, topological spaces can be defined via families of closed sets:

B3 The union of a finite family of closed sets is a closed set.

B4 The intersection of an arbitrary family of closed sets is a closed set.

In topology, as a rule, spaces are considered for which one more axiom holds:

D3 for every x $\theta(x) = x$, where x is the singleton set {x}.

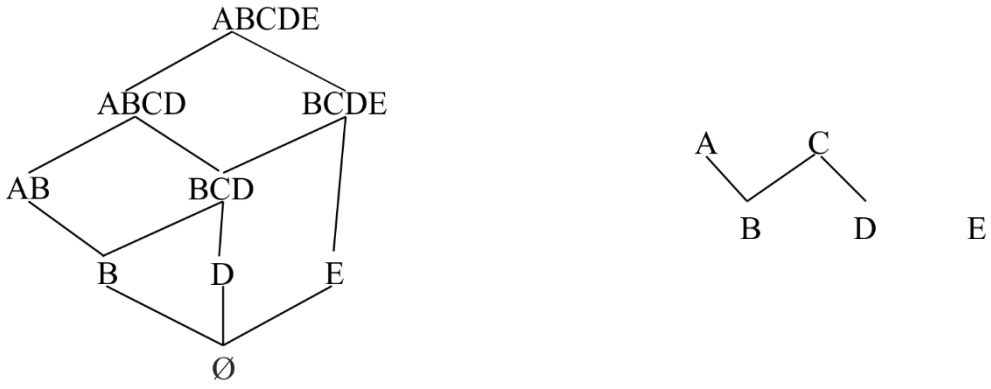
Such spaces are called topological in the strong sense. The requirement of this axiom together with the axiom D1 inevitably turns finite topological spaces into a Boolean U, when any subset of U turns out to be a closed set, so that all the advantages of topology appear only on infinite universes. On the contrary, linguistic spaces are also meaningful at finite universes. In particular, for example, the smallest category in linguistic space is not necessarily an empty set. The fact that spatial representations are also a product of categorization is well known in linguistics for a long time. This comment demonstrates the source of such a connection.

2. *Generators and types.*

Although the linguistic space is a complete lattice, the operations of addition and intersection of categories are not equivalent. We have defined addition by intersection, and this has consequences. In particular, in order to find the result of the intersection of categories, it is enough to know only the operands themselves, and to find the result of addition, one has to use the entire linguistic space. This fact has various manifestations and, in particular, in the generators of the linguistic space, which we will now deal with.

Fig., *a* shows a linguistic space with a universe consisting of five elements U = {A, B, C, D, E}. Categories of space are located in size from top to bottom. The closest in size categories are connected by lines.

As follows from the axiomatics of the categories <B1 and B2>, not all categories are the result of the intersection of some other categories. For Fig., *a* these categories include {ABCD}, {BCDE}, {AB}, {D}, {E}. The category {ABCDE} is not listed here, because it is always included in this list on the basis of B1. Using the selected categories, you can restore the linguistic space by adding all sorts of missing intersections.



a *b*

Linguistic space and hierarchy of its types

Definition 11. The **⋂-generators** of a linguistic space are categories that cannot be represented as intersections of other categories.

Similarly, we can give definitions for other operations.

Definition 12. The **⋃-generators** of a linguistic space are non-empty categories that cannot be represented as a union of other categories.

Definition 13. The **Σ-generators** of a linguistic space are non-empty categories that cannot be represented as the addition of other categories.

For Fig., *a*, the categories {B}, {D}, {E}, {AB}, {BCD} refer to the ⋃-generators, and to the ∑-generators the categories {B}, {D}, {E}, {AB}. ∩-generators and ⋃-generators have significant differences. If the entire linguistic space can be restored from the ∩-generators, then for the ⋃-generator the situation is somewhat different. Unfortunately, not every union of ⋃-generators (as opposed to topological spaces) is a category. So from Fig., *a* it follows that the union of ⋃-generators {AB} and {E} is not a category. Similarly, it is impossible to restore the linguistic space and the ∑-generators. For example, from Fig., *a* it is clear that the category {BCD} cannot be obtained by any addition of ∑-generators without attracting knowledge of the whole space.

Definition 14. A *type* in a linguistic space is a maximal set of elements that fall into the same categories.

Proposition 15. A set types of linguistic space forms a partition of U.

The proof is obvious, since the relation “to enter one and the same type” is reflexive, symmetric and transitive, therefore, it is an equivalence. This in turn means that the classes of this equivalence, dividing U into disjoint sets, are just types.

Proposition 16. Each category of linguistic space consists precisely of the union of certain types.

Proof. Let some type of only partly fall into one category X, and partly into Y. This, however, will contradict Definition 14, since all the elements of a type are grouped by the reason they belong to the same categories.

The investigation. For any two types, there is a separating category that includes one type and does not include the other.

Like Proposition 15, this statement directly follows from the definition of the type: if there was no such separating category, two types would merge into one.

Proposition 17. In each ∪-generator there is one type that is not included in any other category belonging to this ∪-generator. In addition, each type is such a distinguished type for any ∪-generator.

Proof. Let X be a ∪-generator. Therefore, X cannot be obtained by joining smaller categories. Thus, in X there are types that are absent in smaller categories of linguistic space. If there are more than one such types, then, in accordance with the consequence of Proposition 16, there is a category Y, which includes one type and does not include another one. The category Y cannot belong to X and X cannot belong to Y (otherwise Y will not separate types), therefore, Y is incomparable with X. But then $X \cap Y$ is a category belonging to X and containing the type that is not supposed to be contained in categories smaller than X. Contradiction. Therefore, such a type in generator X is unique. We choose an arbitrary type Z and all categories containing Z. The intersection of these categories is a category containing Z, and the smallest for Z in the sense that any smaller category does not contain Z. Therefore, this category cannot be obtained by joining smaller categories and it is by the definition 12 as ∪-generator.

Definition 18. The type defined in Proposition 17 for the ∪-generator is called *own*.

The investigation. There is a one-to-one correspondence between ∪-generators and types.

And from the proof of Proposition 17 it follows that the same type cannot be own for two ∪-generators, as well as the generator cannot have two own types. Thus, ∪-generators and types are in one-to-one correspondence with each other.

Fig., *b* shows the types linguistic space Fig., *a* sorted by the inclusion ∪-generators for which these types are own. Here, each type consists of exactly one element. Of course, this is usually not the case. Five types divide the universe into five disjoint equivalence classes. ∪-generators {B}, {D}, {E} consist exactly of own type {B}, {D}, {E}, ∪-generator {AB} has its own type {A}, ∪-generator of {BCD} – own type {C}.

Ordered types allow you to restore ∪-generators. If X – some ∪-generator and Z is its own type, then X is obtained by adding to the Z all smaller types. From Fig., *b*, for example, it follows that ∪-generator with its own type {C} is obtained by adding to the type {C} the types {B} and {D}.

Remark 19. From the point of view of researching the linguistic spaces themselves, the content of the types does not matter, therefore, without any loss, one-element types can be used, as in our example. In this case, it is said that we turn to the *factor – spaces* for the equivalence of types. Such a transition is reminiscent of axiom D3 for topological spaces, which, however, is much stronger.

Proposition 20. The ∪-generators are categories of form $\theta(\{x\})$.

Proof. The category $\theta(\{x\})$ is the smallest category containing x. This follows directly from A2. Indeed, if $\{x\} \subseteq \theta(Y)$, then $\theta(\{x\}) \subseteq \theta(Y)$. Thus, $\theta(\{x\})$ is a ∪-generator. Let an element x enter into some type, which by virtue of Proposition 17 is own for some ∪-generator Y. The ∪-generator Y by definition 12 is the smallest category containing the type with the element x, therefore, $\theta(\{x\}) = Y$.

Proposition 21. The set of ⋃-generators contains the set of ∑-generators.

Proof. Let X be a ∑-generator. Consequently, for any categories belonging to X, the closure of their union strictly belongs to X. But then, by virtue of A1, the union itself also strictly belongs to X, therefore, X is ⋃-generator.

Proposition 22. When mixing linguistic spaces, a new set of types is formed by the intersection of types of mixed spaces.

Proof. Let x, y belong to one type of mixed space ($(x, y) \in T_3$, where T_3 is the equivalence relation for the types of mixed linguistic space). This means that the elements of this type simultaneously fall into the same categories of mixed space. These categories of mixed space consist of categories of initial spaces and categories representing the missing intersections. It follows that $(x, y) \in T_1$ and $(x, y) \in T_2$, where T_1 and T_2 are equivalence relations for the types of the first and second mixed spaces, respectively. Thus, $(x, y) \in T_1 \cap T_2$ and $T_3 \subseteq T_1 \cap T_2$. Conversely, let $(x, y) \in T_1 \cap T_2$. In accordance with the equivalence property, the equivalence classes of the relation $T_1 \cap T_2$ consist of all possible intersections of the equivalence classes T_1 and the equivalence classes T_2 and, therefore, (x, y) belong to one of these intersections. From the mixed spaces we choose the ⋃-generators X and Y for which these intersecting types are own. A mixed space consists of various intersections of categories of mixed spaces. Consequently, the intersection of these ⋃-generators also belongs to the mixed space; moreover, elements x and y cannot be divided into different types of mixed space due to the fact that all intersections of the categories of smaller chosen generators do not contain these intersecting types. Therefore, x and y will belong to the same type of mixed space and $T_1 \cap T_2 \subseteq T_3$, so $T_1 \cap T_2 = T_3$.

The investigation. When mixing linguistic spaces, the number of ⋃-generators and types does not decrease.

3. Kinds of categorization.

3.1. From general to specific.

This kind of categorization is based on Theorem 5 and is set out in Remark 6. If certain categories are given, then by adding intersections (particular categories), they can always be turned into linguistic space and thereby define the categorization operation. This is exactly what we will do later in the formation of linguistic spaces for verbal categorization.

3.2. From private to general.

Denote through $X \dashrightarrow z$ the product, denoting that "object z is a special case for objects included in X". In other words, everything common for the objects of the set X is also inherent for the object z. In this case, the object z may have some additional features. Any sense can be put into the understanding of the general: common properties, general behavior, common history, common origin, etc. Suppose we are given a lot of products on the U, and X^0 – arbitrary subset of the U. Define $\theta(X^0)$ in the following way. If the left side of any product belongs to the current set X^i, then the object from the right side is added to X^i and the transition to X^{i+1} is made. We act this way until nothing new can be added to the current set X^j. Obviously, this process is finite, if the number of productions are finite as well.

As an example, we choose a set of five elements U = {A, B, C, D, E}, and define six products: $\{A\} \dashrightarrow \{B\}$, $\{B, D\} \dashrightarrow \{C\}$, $\{C\} \dashrightarrow \{B\}$, $\{C\} \dashrightarrow \{D\}$, $\{D, E\} \dashrightarrow \{C\}$ and $\{B, E\} \dashrightarrow \{C\}$. Let, for example, $X^0 = \{A, E\}$. With the help of the first production we get $X^1 = \{A, B, E\}$. Using the sixth product, we get $X^2 = \{A, B, C, E\}$. Finally, the fourth production process completes $X^3 = \{A, B, C, D, E\} = U$. All the sets constructed in a similar way are shown in Fig. *a*. Essentially, using the

product to the original set of examples particular cases are added, forming, in the end, a category that contains all the examples having the same overall as the original examples.

Theorem 23. The operation θ defined by production is the categorization operation.

Proof. By definition of the algorithm we have $\theta(U) = U$. Suppose θ is not a categorization operation. Then there are two sets of X and Y, obtained by the above procedure such that $X \cap Y \subset \theta(X \cap Y)$. But this means that there is a production of the $W \dashrightarrow z$ such that the $W \subseteq X \cap Y$ but z does not belong to the $X \cap Y$. Since $W \subseteq X \cap Y$, then $W \subseteq X$ and $W \subseteq Y$, hence, in the process of obtaining sets of X and Y should have been added, and z but then $\{z\} \subseteq X \cap Y$. Contradiction. Therefore, θ is a categorization operation.

3.3. Verbal categorization.

Consider the simplest case of verb categorization, where the verb is a binary relation. This categorization is based on observations of the interaction of pairs of objects and is a development of Galois correspondence [3]. The binary case is not difficult to generalize to the relationship of any arity, but this is the subject of a separate conversation. The choice as an example of a binary case is due to the fact that in the overwhelming number of cases the verbs of the language are binary, that is, they correspond to the action between the subject and the object.

So, let $(U \times V, S)$ be the binary relation given on the Cartesian product of the universums U and V, S be the graph of the binary relation. Let $x \in U$. Let x^{Δ} be the set of all elements of $y \in V$ for which $(x, y) \in S$. Similarly, the symbol y^{∇} $(y \in V)$ denotes the set of elements $x \in U$, for which $(x,y) \in S$. Using a categorization of "from general to particular" (subsection 3.1), if we add all possible intersections to sets of the form of x^{Δ} (of the form y^{∇}), then we get some linguistic space. In fact, this means that we extend the operators Δ and ∇ already by multi-element subsets from U and V:

$$X^{\Delta} = \bigcap_{x \in X} x^{\Delta} \ (X \subseteq U); \ Y^{\nabla} = \bigcap_{y \in Y} y^{\nabla} \ (Y \subseteq V).$$

For an empty subset, we set $\varnothing^{\Delta} = V$; $\varnothing^{\nabla} = U$. Thus, for each $X \subseteq U$ and $Y \subseteq V$ we associate a certain category from the linguistic space on V and U, respectively. Consider the properties of these spaces and the relationship between them.

Proposition 24. If $X_1 \subseteq X_2$ $(Y_1 \subseteq Y_2)$, then $X_2^{\Delta} \subseteq X_1^{\Delta}$ $(Y_2^{\nabla} \subseteq Y_1^{\nabla})$.

Proof. Since X1 ⊆ X2, the set of intersecting sets for X_2^{Δ} includes all intersecting sets for X_1^{Δ}, and hence $X_2^{\Delta} \subseteq X_1^{\Delta}$. Similarly for Y_1 and Y_2.

The investigation. Let $X_1 \subseteq X_2$ $(Y_1 \subseteq Y_2)$, then $X_1^{\Delta\nabla} \subseteq X_2^{\Delta\nabla}$ $(Y_1^{\nabla\Delta} \subseteq Y_2^{\nabla\Delta})$.

Proposition 25. For any X (Y), $X \subseteq X^{\Delta\nabla}$ $(Y \subseteq Y^{\nabla\Delta})$ holds.

Proof. By the definition of the empty set $\varnothing \subseteq \varnothing^{\Delta\nabla}$ $(\varnothing \subseteq \varnothing^{\nabla\Delta})$. If $X \neq \varnothing$, but the intersection of X^{Δ} is empty, then $X^{\Delta\nabla} = \varnothing^{\nabla} = U$ и $X \subseteq X^{\Delta\nabla}$. Let $X^{\Delta} \neq \varnothing$. Then, for any $y \in X^{\Delta}$ occurs $(x,y) \in S$ for every x from X. It follows that the intersection of all y^{∇} for $y \in X^{\Delta}$ contains X, therefore $X \subseteq X^{\Delta\nabla}$. Similarly, for Y.

Proposition 26. For any X_1, X_2 (Y_1, Y_2) is true $(X_1 \cup X_2)^{\Delta} = X_1^{\Delta} \cap X_2^{\Delta}$ $((Y_1 \cup Y_2)^{\nabla} = Y_1^{\nabla} \cap Y_2^{\nabla})$.

The proof follows directly from the associativity of the intersection.

Proposition 27. $X^{\Delta\nabla} = X^{\Delta\nabla\Delta\nabla}$ $(Y^{\nabla\Delta} = Y^{\nabla\Delta\nabla\Delta})$.

Proof. From Proposition 25 and the investigation to Proposition 24, we have $X^{\Delta\nabla} \subseteq X^{\Delta\nabla\Delta\nabla}$. On the other hand, by virtue of Proposition 25 $X^{\Delta} \subseteq (X^{\Delta})^{\nabla\Delta}$, therefore, taking into account Proposition 24 $(X^{\Delta})^{\nabla\Delta\nabla} \subseteq X^{\Delta\nabla}$, which means $X^{\Delta\nabla} = X^{\Delta\nabla\Delta\nabla}$. Similarly, for Y.

Now note that proposition 25 proves the validity of axioms A1, a consequence of Proposition 24 – validity C2 and Proposition 27 – validity C1. Thus, the following

Theorem 28. The operations $\Delta\nabla$ ($\nabla\Delta$) are categorization operations on U (V).

The investigation. By virtue of Lemma 3

$$X_1^{\Delta\nabla} \cap X_2^{\Delta\nabla} = (X_1^{\Delta\nabla} \cap X_2^{\Delta\nabla})^{\Delta\nabla} \qquad (Y_1^{\nabla\Delta} \cap Y_2^{\nabla\Delta} = (Y_1^{\nabla\Delta} \cap Y_2^{\nabla\Delta})^{\nabla\Delta}).$$

Thus, the operations $\Delta\nabla$ ($\nabla\Delta$), generated by the relation S (verb), satisfy the categorization axioms, therefore we will call them a verb categorization. When at the beginning of this subsection we used the categorization of "from general to particular", it meant that the operations Δ and ∇ are, by definition, build linguistic spaces on the universes U and V respectively. But the operation $\Delta\nabla$ ($\nabla\Delta$) is quite different and therefore theorem 28 is not obvious. The following concludes this discussion.

Proposition 29. The linguistic spaces in the operations $\nabla\Delta$ and Δ ($\Delta\nabla$ and ∇) coincide.

Proof. By Proposition 25 $X^{\Delta} \subseteq (X^{\Delta})^{\nabla\Delta}$ и $X \subseteq X^{\Delta\nabla}$. Hence, with regard to the proposition 24 $(X^{\Delta\nabla})^{\Delta} \subseteq X^{\Delta}$, and therefore, $(X^{\Delta\nabla})^{\Delta} = X^{\Delta}$, and each Δ-category included linguistic space operations $\nabla\Delta$. Conversely, each category Y by operation $\nabla\Delta$ is also a category $(Y^{\nabla})^{\Delta}$ by operation Δ. Therefore, the operations $\nabla\Delta$ and Δ ($\Delta\nabla$ and ∇) coincide and form the same linguistic spaces.

So, to obtain a linguistic space on U (V), one can either use sets X, categorizing them with the operator $\Delta\nabla$, or sets Y, translating them by mapping ∇ into categories of the same linguistic space.

Definition 30. In the following, we will call the linguistic space on U simply a *space*, and the linguistic space on V as a *co-space*.

Definition 31. The mapping Δ between the categories of space and co-space will be called the *verb* (binary verb), and the map ∇ between the categories of co-space and space is called the *co-verb* (reflexive verb).

Lemma 32. For X_1, X_2 (Y_1, Y_2) inclusion of $X_1^{\Delta} \subseteq X_2^{\Delta}$ ($Y_1^{\nabla} \subseteq Y_2^{\nabla}$) is satisfied if and only if $X_2^{\Delta\nabla} \subseteq X_1^{\Delta\nabla}$ ($Y_1^{\nabla\Delta} \subseteq Y_2^{\nabla\Delta}$).

Proof. Let $X_1^{\Delta} \subseteq X_2^{\Delta}$, then by virtue of Proposition 24 $X_2^{\Delta\nabla} \subseteq X_1^{\Delta\nabla}$. Conversely, let $X_2^{\Delta\nabla} \subseteq X_1^{\Delta\nabla}$, then by virtue of Proposition 24 $X_1^{\Delta\nabla\Delta} \subseteq X_2^{\Delta\nabla\Delta}$ and taking idempotency into account, we get $X_1^{\Delta} \subseteq X_2^{\Delta}$. Similarly, for Y.

In fact, the verbs Δ and ∇ are the same mapping (direct and inverse), which turns over inclusions and operations in space (providing dualism).

Theorem 33. The space is dual isomorphic to the co-space by the verb Δ in the sense that under isomorphism the inclusion relation changes direction, the intersection operation goes into addition operation, and the addition operation goes into intersection operation. Similar wording for the co-verb.

Proof. We show that the verb is a bijection, moreover, $\Delta^{-1} = \nabla$. For each category of X, we have $(X^{\Delta})^{\nabla} = X$. Thus, $\Delta\nabla$ is the identity mapping (similarly for $\nabla\Delta$). So, the number of categories in the space and to co-space is alike. Lemma 32 implies a change in the direction of inclusion when moving along the verb and the co-verb.

For X_1 , X_2 in accordance with Proposition 26, we have:

$$(X_1+X_2)^{\Delta} = ((X_1 \cup X_2)^{\Delta\nabla})^{\Delta} = (X_1 \cup X_2)^{\Delta} = X_1^{\Delta} \cap X_2^{\Delta}.$$

$$\text{Further, } (X_1 \cap X_2)^{\Delta} = ((X_1^{\Delta})^{\nabla} \cap (X_2^{\Delta})^{\nabla})^{\Delta} = (X_1^{\Delta} \cup X_2^{\Delta})^{\Delta\nabla} = (X_1^{\Delta} + X_2^{\Delta}).$$

We defined the categorization axioms, and also considered its three versions and some properties that will be needed later. This does not mean that there are no other categorization options. It is possible that readers will offer some other option. The main thing is that this process satisfies the formulated axioms.

Of the considered categorization options, the main one, of course, is the verb. After the verb categorization has been completed, it is possible to build a system for identifying categories by signs, after which it becomes a model. Namely, when confronted with an object after its identification into a category, it is possible to predict its behavior using verbs.

Of course, after categorization, the world does not always give us signs that allow us to identify categories. What appears to be an anomaly in cognitive linguistics looks completely natural in R-linguistics. A well – known example is Ludwig Wittgenstein with the category "game" for which there is no definition in the signs. The category "game" is determined not by signs, but by the action "play". Similarly, the category "predators" includes such objects, such as the sundew, mantis, perch and lion. These objects have no common signs. They are united by the verb "eat", according to which the objects of this category are eaten by the flesh of other creatures. Often the verbal categorization is hidden and the categorization is given the process of building an identification system. Let's say that Karl Linney was looking for suitable attributes when building his classification system for animals. What allowed him to choose some signs and refuse others? In fact, even before the selection of signs, there was already some categorization in his head, which could later be clarified. Most likely, his categorization was based on the relationship of "kinship". Problems Linnaeus were not associated with categorization, and with identification. So, the problems with the signs made Linnaeus in determining the primates to choose characteristics "five fingers" and "two mammary glands", after which the primates have become sloths and bats. We consider this choice unsuccessful precisely because our reason for categorizing animals also uses the relationship of "kinship", although it is already more perfect.

It is natural to assume that since the "predators" category does not have any clear identifying features, the boundaries of this category are not defined and it is blurred or fuzzy. Here I want to draw attention to the fact that the axioms categorization A1 and A2 do not contain any pre parcels to the fuzziness. After all, you cannot fuzzily attack another creature and fuzzily swallow its flesh. Each of us also quite clearly determines that he has fallen ill (has moved from the category of healthy people to the category of patients) by changes in his behavior: loss of appetite, shortness of breath, lameness, etc. We are quite clearly going to the doctor, because we clearly record the fact of a change in our behavior. The discussion of fuzzy categorization is of course a separate topic, and I touch it now, only to draw the reader's attention to the relationship between the idea of fuzzy categorization and the R-linguistics approach.

Not all relationships into which objects enter generate categories, because far from all relationships are stable. For example, the relationship of "friendship" in the universe of people is not sustainable. It changes with time and does not affect the external signs of the person himself. Or the "reading" attitude that people enter into and the books are also constantly changing. There are a huge number of verbs that use categories already formed by other verbs (relations). Say, not only a person or a dog can go, but some process or time. Usually the verbs that make up the categories appear to be fairly stable relations, which, as a rule, form some signs reflecting the fact of these permanent relations. For example, the relationship between animal species and habitat regions is reflected in animals by the appearance of certain traits. Animals living in the far north have either thick fur or a thick layer of fat. Typically, such stable relationships manifest themselves in data dependencies, which allow them to be extracted from complex linguistic structures. This is also a separate topic for discussion, stemming from the categorization.

It should be noted that in the surrounding world there is not only a binary interaction. Surprisingly, the more complex interaction also generates categories with the axiomatics described here. It is also a matter for future discussions.

Finally, it must be said that the object and subject of the action can be not only categories, but also variables, the simplest of which are pronouns. And this is only a very small part of the concepts that are widely used in the language, but are not categories.

Conclusion. The content of this article is not quite traditional for linguists, and the reader should answer the following question: "Does the definition of the concept in language belong to linguistics?" If you answer this question in the affirmative form and cannot object to the axiomatics given, you will have to admit that all the results following the axioms are directly related to linguistics. You will be surprised how long this approach does not require any reference to the language or even to the existence of the world. This is a good sign given the linguistic diversity around us. Although we are only at the beginning of this journey, several important observations and conclusions can be drawn.

REFERENCES

1. Dunaev, V.V. and Polyakov, O.M. (1987), "Methodological aspects of relational classification theory", *NTI*, ser. 2, no. 4, pp. 21–27.

2. Moore, E. (1910), "Introduction to a form of general analysis", *AMS Colloquium Publishing*, Yale Univ. Press, New Haven, USA.

3. Birkhoff, G. (1984), *Teoriya reshetok* [Lattice Theory], Transl. by Salii, V.N., Nauka, Moscow, Russia.

4. Polyakov, O. M. (1986), "Klassifikatsionnaya model' dannykh", *NTI*, ser. 2, pp. 13–20.

5. Kuratowski, K. (1966), *Topologiya* [Topology], Transl. by Antonovskii, M.Ya., vol. 1, Mir, Moscow, Russia.

Information about the author.

Polyakov Oleg Maratovich – Can. Sci. (Engineering) (1982), Associate professor at the Department of Information Technology of Entrepreneurship, Saint-Petersburg State University of Aerospace Instrumentation, 67 lit. A, Bol'shaya Morskaya str., St Petersburg 190000, Russia. The author of 30 scientific publications. Areas of expertise: linguistics, artificial intelligence, mathematics, database design theory, philosophy. E-mail: road.dust.spb@gmail.com

СПИСОК ЛИТЕРАТУРЫ

1. Дунаев В. В., Поляков О. М. Методологические аспекты реляционной теории классификации // НТИ. Сер. 2. 1987. № 4. С. 21–27.

2. Moore E. Introduction to a form of general analysis // AMS Colloquium Publishing. New Haven: Yale Univ. Press, 1910.

3. Биркгоф Г. Теория решеток / пер. В. Н. Салий. М.: Наука, 1984.

4. Поляков О. М. Классификационная модель данных // НТИ. Сер. 2. 1986. № 9. С. 13–20.

5. Куратовский К. Топология / пер. М. Я. Антоновского. М.: Мир, 1966. Т. 1.

Информация об авторе.

Поляков Олег Маратович – кандидат технических наук (1982), доцент кафедры информационных технологий предпринимательства Санкт-Петербургского государственного университета аэрокосмического приборостроения, ул. Большая Морская, д. 67, лит. А, Санкт-Петербург, 190000, Россия. Автор более 30 научных публикаций. Сфера научных интересов: лингвистика, искусственных интеллект, математика, теория проектирования баз данных, философия. E-mail: road.dust.spb@gmail.com