Языкознание
Linguistics

УДК 81–116 http://doi.org/10.32603/2412-8562-2020-6-2-107-114
Оригинальная статья / Original paper

Linguistic Data Model for Natural Languages and Artificial Intelligence. Part 4. Language

Oleg M. Polyakov✉

Sait-Petersburg State University of Aerospace Instrumentation, St Petersburg, Russia

✉road.dust.spb@gmail.com

Introduction. The paper continues a series of publications on linguistics of relations (hereinafter R–linguistics) and is devoted to questions of the formation of a language from a linguistic model of the world. Moreover, the language is considered in its most general form, without taking into account the grammatical component. This allows you to focus on the general problems of language formation. Namely, this allows us to show why language adequately reflects the model of the world and what are the features of the transition from model to language. This new approach to language is relevant in connection with the formation of an understanding of the common core in all natural languages, as well as in connection with the needs for the formation of artificial intelligence subsystems of interaction with humans.

Methodology and sources. Research methods consist in the formulation and proof of theorems about language spaces and their properties. The materials of the paper and the given proofs are based on the previously stated ideas about linguistic spaces and their decompositions into signs.

Results and discussion. The paper shows how, in the most general form, the formation of language structures takes place. Namely, why does language adequately reflect the linguistic model, and what is the difference between linguistic and language spaces? The concepts of an open and closed form of the language are formulated, as well as the law of form. Examples of open and closed forms of the language are shown. It is shown that the formation of the language allows you to compensate for the lack of real signs in the surrounding world while maintaining the prognostic properties of the model.

Conclusion. Any natural language is a reflection of the human world model. Moreover, all natural languages are similar in terms of the principles of forming the core of the language (language space). Language spaces standardize the models of the world by equalizing real and fictional signs of categories. In addition, the transition to language simplifies some of the problems of pattern recognition and opens the way to the logic of natural language.

Key words: R–linguistics, language, open form, closed form, generators.

For citation: Polyakov O. M. Linguistic Data Model for Natural Languages and Artificial Intelligence. Part 4. Language. DISCOURSE. 2020, vol. 6, no. 2, pp. 107–114. DOI: 10.32603/2412-8562-2020-6-2-107-114

Conflict of interest. No conflicts of interest related to this publication were reported.

Received 23.01.2020; adopted after review 27.02.2020; published online 27.04.2020

Лингвистическая модель данных для естественных языков и искусственного интеллекта. Часть 4. Язык

О. М. Поляков✉

Санкт-Петербургский государственный университет аэрокосмического приборостроения, Санкт-Петербург, Россия

✉road.dust.spb@gmail.com

Введение. Статья продолжает серию публикаций по лингвистике отношений (далее R–лингвистика) и посвящена вопросам формирования языка из лингвистической модели мира. При этом язык рассматривается в самом общем виде без учета грамматической составляющей. Это позволяет сосредоточиться на общих проблемах формирования языков. А именно показать, почему язык адекватно отражает модель мира и в чем особенности перехода от модели к языку. Это новый подход является актуальным в связи с формированием понимания общего ядра во всех естественных языках, а также в связи с потребностями формирования для искусственного интеллекта подсистем взаимодействия с человеком.

Методология и источники. Методы исследования заключаются в формулировке и доказательстве теорем о языковых пространствах и их свойствах. Материалы статьи и приведенные доказательства базируются на изложенных ранее представлениях о лингвистических пространствах и их разложениях в признаки.

Результаты и обсуждение. Показано, как в самом общем виде происходит формирование языковых структур. А именно, почему язык адекватно отражает лингвистическую модель, и в чем отличие языковых и лингвистических пространств. Сформулированы понятия открытой и закрытой форм языка (приведены примеры), а также закон формы. Продемонстрировано, что формирование языка позволяет компенсировать недостаток в окружающем мире реальных признаков с сохранением прогностических свойств модели.

Заключение. Любой естественный язык является отражением модели мира человека. При этом все естественные языки схожи в части принципов формирования ядра языка (языкового пространства). Языковые пространства стандартизуют модели мира тем, что уравнивают реальные и вымышленные признаки категорий. Кроме того, переход к языку упрощает некоторые проблемы распознавания образов и открывает дорогу к логике естественного языка.

Ключевые слова: R–лингвистика, язык, открытая форма, закрытая форма, образующие.

Для цитирования: Поляков О. М. Лингвистическая модель данных для естественных языков и искусственного интеллекта. Часть 4. Язык // ДИСКУРС. 2020. Т. 6, № 2. С. 107–114. DOI: 10.32603/2412-8562-2020-6-2-107-114

Конфликт интересов. О конфликте интересов, связанном с данной статьей, не сообщалось.

Поступила 23.01.2020; принята после рецензирования 27.02.2020; опубликована онлайн 27.04.2020

Introduction. In terms of programmers, language is a means of exporting / importing data in a linguistic model of the world from one person to another. This involves inventing an entire coding system that allows model data to be packaged so that the other end can unpack the data and restore the corresponding structures or use them. It is clear that the problem of export/import arises only insofar as there are models themselves. The solution of the export/import problem is determined by both the nature of the model and the ingenuity of the team of programmers, its previous practices, as well as good ideas borrowed from neighbors.

So, in the language there is a fixed part associated with the model, as well as a variable part associated with the implementation of specific ideas for solving the problem of export/import by a specific team. The second changing part is usually called grammar. This is a set of solutions that allow you to invest in a sequence of symbols (sounds, gestures, etc.) elements of linguistic models so that on the receiving side of this sequence it can be decompressed and restore the model. From the point of view of R-linguistics, it is a particular problem solved by each group of native speakers in their own way. This does not mean that grammar is an unimportant secondary field. Without knowledge of these rules, it is impossible to translate and solve many other problems. But the study of these sophisticated solutions is not within the scope of R-linguistics: the subject of its interest is part of the problem associated with the linguistic model itself.

As an example, consider the export/import of Excel tables. Spreadsheets define some principle of organization of the computational process, and each table is a specific model of some process. If we want to transfer tables, somehow it is necessary to transfer cell numbers consisting of two variables, data types in cells, functions, references to other cells, etc. Other words, the nature of the transferred structures will certainly determine the data set and their relationships to be transferred. Actually, this is what R-linguistics is interested in. But the task of "pulling" a spreadsheet into a temporal sequence of the communication channel can be solved in thousands of different ways, which actually determine the specific "grammar" of the model export/import. This is a private task solved by each group of native speakers in their own way.

As a research method, the results obtained in the previous articles of the series are used, which are interpreted and analyzed here. On the basis of this, the mathematical foundations of language formation and the difference between language spaces and linguistic ones are formulated.

Results and discussion.

The emergence of language.

Everything that we have considered so far fits into the following scheme. Modeling the observed relationships leads us to the dualism of the subject and addition through verbs. To take advantage of this dualism and turn it into a means of prediction, it is supplemented by the dualism of nouns and adjectives (signs).

But one day one consciousness, based on theorem 5 [1], was looking for a sign that is included in the only irreducible representation of some linguistic space. Whether this consciousness was lazy, or dull, or indeed the world had clearly failed him, this sign could not be found. And then consciousness decided to postpone this search for later, and in the meantime, in order to be able to think and make predictions, replace the result of the analysis of this attribute simply with some symbol, since there is no difference between this value and any other label.

This story is not as fantastic as it might seem at first glance, although it is not a historical plot. The irreducible representation in Theorem 5 consists of the simplest spaces, each of which includes a universe and an ∩-generator. Each element of the decomposable space is then obtained by the intersection of all elements from the spaces included in the decomposition and containing this element. So, we can mark up the linguistic space with the names of nonzero values of the calculated signs. Namely, we associate each ∩-generator with its own symbol (the name of a nonzero value of the corresponding sign). The name of any category of space will consist of a collection of category names containing this category. For example, the universe will receive the name $И_0$, the nearest ∩-generator will receive the name $И_0И_1$, and so on.

Example 1. Consider the space from Example 11 [1]. In fig. 4*c* [1], we replace the contents of the categories on the constructed names. The result of such a replacement is shown in fig. Here, for convenience, an empty set does not have a name and a verb, since it can have any names (it is the same in all spaces) and the verbs from it act on all universes. If we now replace the formal symbols $И_0$ with FISH, $И_1$ with PREDATORS, $И_2$ with VICTIMS, $И_3$ with LARGE, we get the following statements: LARGE FISH PREDATORS eat FISH, FISH PREDATORS eat FISH VICTIMS, etc.

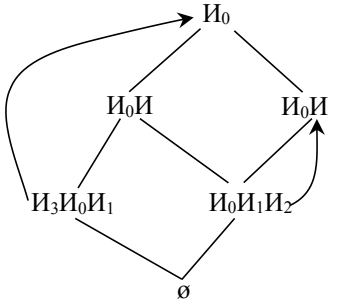

Fig. Language namespace

There's no recognition here yet, so there's no prediction. For example, the name "FISH PREDATORS" cannot yet be calculated from real fish, and it is actually a label. If, as the model develops, a system of real signs and parameters is formed to assess their values, the values of the signs can be taken into account in the names and thus the model will receive a prognostic character. For example, replacing the name PREDATORS with TOOTHY will give the following statement: LARGE TOOTHY FISH EAT FISH. Here already there is forecast, because, having caught walleye, assessing his the size of the and discovering have him her teeth, can be predict his aggression relative to all fish.

Definition 2. The category name structure of a linguistic space is called a language space. Names are treated simply as sets of words.

By construction, it does not matter for the language space what is used as names: some labels or names of values of real signs. Let α be a mapping that maps each category of linguistic space $\mathbb{P}$ to its name.

Theorem 3. Linguistic and language spaces are dually isomorphic.

Proof. By construction, the mapping α is surjective. If for any categories X, Y of $\mathbb{P}$ turns out to be $\alpha(X) = \alpha(Y)$, it will mean that X and Y are in the same ∩-generators, and therefore $X = Y$ and α – bijection. If $X \subseteq Y$, then X belongs to all those ∩-generators to which Y belongs, and therefore $\alpha(Y) \subseteq \alpha(X)$. Conversely, if $\alpha(Y) \subseteq \alpha(X)$, then X belongs to all those ∩-generators as Y, hence $X \subseteq Y$.

We show that $Z = X + Y$ is equivalent to $\alpha(Z) = \alpha(X) \cap \alpha(Y)$. Let $Z = X + Y$. By virtue of the dual isotonicity of the map, we have $\alpha(Z) \subseteq \alpha(X)$, $\alpha(Z) \subseteq \alpha(Y)$, and $\alpha(Z) \subseteq \alpha(X) \cap \alpha(Y)$. Back, $\alpha(X) \cap \alpha(Y)$ includes the names of all ∩-generators containing both X and Y, and therefore $X \cup Y$, and, respectively, Z. Hence $\alpha(X) \cap \alpha(Y) \subseteq \alpha(Z)$ and $\alpha(Z) = \alpha(X) \cap \alpha(Y)$. Let now $\alpha(Z) = \alpha(X) \cap \alpha(Y)$. Hence $X \subseteq Z$, $Y \subseteq Z$, $X \cup Y \subseteq Z$ and, by definition of the addition operation $X + Y \subseteq Z$. Let there be a category T such that $X + Y \subseteq T \subset Z$. Therefore, in the representation T there is an ∩-generator that is not in the representation Z, therefore, this generator is also in the representation $X + Y$, and therefore both in $\alpha(X)$ and $\alpha(Y)$. But this contradicts the assumption $\alpha(Z) = \alpha(X) \cap \alpha(Y)$.

We show now that $Z = X \cap Y$ is equivalent to $\alpha(Z) = \alpha(X) + \alpha(Y)$. Let $Z = X \cap Y$ be true, then Z is the result of the intersection of all ∩-generators for X and Y. In other words, $\alpha(Z) = \alpha(X) \cup \alpha(Y)$. But α is a bijection, and if $\alpha(X) \cup \alpha(Y)$ is not a closed set, then α maps Z to the nearest closed set containing $\alpha(X) \cup \alpha(Y)$, that is, to the set $\alpha(X) + \alpha(Y)$. Conversely, let $\alpha(Z) = \alpha(X) + \alpha(Y)$ be true. By the definition of the addition operation, $\alpha(X) \cup \alpha(Y) \subseteq \alpha(X) + \alpha(Y)$ and therefore $Z \subseteq X \cap Y$. If $Z \subset X \cap Y$, then there is an ∩-generator of T such that T is present in the representation Z, but it does not exist in the representations X and Y. But then $\alpha(X) + \alpha(Y) \subset \alpha(Z)$, which contradicts assumption. Thus, $Z = X \cap Y$.

Remark 4. Theorem 3 states that a language space is actually a co-space on the verb α. There is nothing surprising. The language should preserve all the information about the constructed linguistic model. All this would be just a side view for the co-spaces already considered in the first part, if not for the next.

Theorem 5. In the language space, ∪-generators coincide with ∑-generators.

Proof. According to proposition 21 [2] it is only necessary to prove that there are no categories among ∪-generators that are not included in the set of ∑-generators. So let $\alpha(Z) = \alpha(X_1) + ... + \alpha(X_n)$ (α (Z) is not part of ∑-generators), where $\alpha(X_1),..., \alpha(X_n)$ are ∑-generators of the language space, and let α (Z) be ∪-generators. This means that there is a name I of its own type, which is not included in any set of names $\alpha(X_1),..., \alpha(X_n)$. According to theorem 3, $Z = \bigcap_{i=1}, ..., {}_n X_i$, where $X_1, ..., X_n$ are ∩-generators. Hence on the rule for constructing names in α(Z) may not be a name not occurring in $\alpha(X_1),..., \alpha(X_n)$, and hence α(Z) is ∪-forming.

Theorem 6. Linguistic space is language if and only if each ∪-generator has no more than one ancestor.

Proof. Let the linguistic space is not language space. Therefore, its ∪-generators do not coincide with the ∑-generators. By Proposition 21 [2], there exists a ∪-generator X, which is not a ∑-generator. This means that the ∪-generator of X is the sum of at least some two ∪-generators of Y and Z, so that the ∪-generator of X has at least two ancestors. Conversely, let ∪-generators and ∑-generators coincide in the linguistic space. Therefore, any ∪-generator is simultaneously a ∑-generator and, therefore, it cannot be obtained by adding any other ∪-generators. This means that no ∪-generator is the sum of the other ∪-generators. Therefore, any ∪-generator has at most one ancestor.

The investigation. Having one ancestor for each ∪-generator means that each ∪-generator coinciding with its own type is the root of a tree in the species hierarchy. In other words, we are dealing with a forest (a set of trees by the number of such ∪-generators). Theorem 6 ensures that no two trees in this forest touch each other. This is a very good property when it comes to types recognition [3]. The types recognition system for language spaces can be constructed as follows. First, the recognition of forest root types (minimum types) is built. Let some type be characterized by some values of k parameters. Let's consider some of its daughter type. The preceding type details the daughter type. Therefore, the daughter type is characterized only by some part of the values of k parameters. Thus, we only need to determine for each daughter types what parameters can be discarded when recognizing it.

In addition, this property of language spaces opens the way to logic, because it ensures the implementation of the law of exclusion of the third. I will postpone the discussion of this thesis until the next part, where the logic of natural language will be discussed.

The law of form.

In [3], when discussing the problem of recognition, we have already considered what happens when the calculated signs cannot be found in the world around us. The same problem arises in language, but it is of a different nature.

In order to determine the category in question in the language, it is necessary to specify the signs that characterize it. We literally have to pronounce all the values of signs that could not find a physical analogue. We will call it an open form of language, because it is forced to reflect the structure of the language space through the announcement (written or oral) of the meanings of signs. For example, the sign of the number of protons in the nucleus of an atom is a difficult sign to measure, so the speaker has to add the values of replacement signs: alkaline metals, rare earth metals, etc. Sometimes it is not possible to find only a part of the signs and the values of the signs are made public only from this part. Let's say, in the example with metals, the identification of the category "metal" in the signs is not disclosed.

The described situation characterizes the most important problem of language and thinking in general. This is a closed form of the language, which is understood as the removal from the language of signs characterizing a category and its place in the linguistic space. When we pronounce the word "metal", we don't tell the listener by what signs we define this category, because these signs are enough to identify the metal. We believe that the listener uses exactly the same signs, and there is no need to tell him about it. Yes, indeed, many signs have a socio-historical origin and most likely they are the same in different people. This is also facilitated by dictionaries in which, regardless of the "physical" existence of signs, the definition of categories is given in a literal indication of the signs and their meanings. However, in general, we do not know what signs characterize the category of the interlocutor. If I tell you now that bamboo belongs to the cereal family, many readers will probably be surprised, and only this will allow me to suspect that the "size" sign is on their list of sign for the category of cereal plants.

Of course, this is the reason for all kinds of communicative failures, since the part of the linguistic model transmitted through the language and the part that is formed under the influence of the language on the receiving side can differ significantly. For the same reason, most of the names of categories in the process of communication are in fact just labels, denoting a hidden conjunction from language meanings of signs that a person determines subconsciously and does not pronounce. For example, none of us will say, "I saw a chordal cranial cloven-hoofed mammal milked today". Everyone will say, "I saw a cow milked today". The cow has enough real signs to identify it and we do not transmit this data in the process of communication. In General, this situation can be formulated in the form of the law of form*: the worse the situation is with the parameters for recognizing the values of signs, the more the language tends to open form, and vice versa***.**

In other words, the worse the external data is, the more we have to talk. This problem is one of the most important stimulus for the emergence of language, which I jokingly described at the beginning of the article. The more deeply we know the world, the more complex the model of the world becomes, and the more often we are faced with a lack of parameters. We have to invent signs like "strangeness" or "fascination" in quantum physics that we can't at first calculate or even describe. The development of the world model forces the appearance of language primarily in an open form. This is not a communicative reason for the emergence of a language, but cognitive not in the sense of the way of preserving knowledge, but in the sense of the process of

cognition. And this does not mean that we know the world through language. This means that without language we cannot standardize (unify) a sufficiently complex model in our consciousness.

Conclusion. In conclusion, consider an example. From the above it may seem that in its pure form, the open form of the language is some kind of exotic, little suitable for forecasting. This is not true. Consider an example with UDC symbols for identifying books. So there are many people and many books. People read books and thereby form a relation READING. Of course, there is not a single person who knows the whole relation, but there are people, bibliographers who have a good idea of him. The main task of the bibliographer is to create, based on this knowledge, a system of predicting which books to offer to a person when he comes to the library or bookstore, or which books (manuscripts) to purchase at a book exhibition (from the authors) to satisfy readers, etc. It is clear that a person, as a rule, will not read a book already read. Therefore, predictions should relate to reading that has not yet been realized. In the framework of the relational model, this is fundamentally impossible to do.

So, on the basis of the existing READING relation, it is necessary to build space on set readers and co-space on set books. After this, it is necessary to build categories recognition systems for people and for books. The relation READING is variable and does not leave a mark on people of what they read, although perhaps an experienced librarian can form an idea of the client's reading interests by mere appearance, behavior and speech. Here, however, we are lucky. People perfectly identify themselves as a type of reader. For example, I say inside myself: ‘‘Today is a good mood. I haven't read anecdotes in a long time. Today I am a humorist reader’’. Tomorrow this reader may already be a dentist or a topological mathematician. In short, the reader identifies himself and reports the results of this identification to the librarian when he answers a question about what interests him.

Now we need to build a book recognition system. First, the parameters are searched and their values are determined for recognition. This is a system of keywords used in books. You can use a certain technique to find these words and determine which readers will read this particular book. From now on, the linguistic model is ready. Now you can group books on the shelves according to their occurrence in certain categories. This distribution is determined by keywords. Great? Yes, but not really. After all, in accordance with the wishes of the reader, it is necessary to let him into the repository of books to those sections that are formed. Therefore, they act otherwise.

Categories implemented as a sets of books grouped on shelves are replaced by the names of these categories with the special structure described in this article. These names are constructed so that they accurately reflect the linguistic space of books resulting from the READING relationship: a large number of readers have few common books to read, and Vice versa. So it is with names: books of General content that have many readers will get names that consist of a small number of words, and books of special content that have a limited number of readers will get names that consist of a large number of words. You can use regular decimal digits separated by a period as words. This is how the universal decimal classification (or UDC codes) is obtained.

Now you can move away from the real world of books by creating a catalog – a prototype of the language model, but only for books and only for one relation. After that, you can run people there with their self-identification, and let them do their own forecasts.

The result is an open language, but sentences in this language are partially hidden. More precisely, here, due to the specifics, there is always one typical proposal. The subject is the name

of the reader's self-identification. It is in his head and is not pronounced, since he himself is the consumer of this proposal. The same verb "read" always acts as a verb, therefore it is also not pronounced. The UDC code or its verbal decoding acts as an addition. This language demonstrates the amazing resourcefulness of man: language is created for a variable relation; one side of objects is not only not identified by signs, but depending on its state changes its behavior in this relation. The relation itself is known only by parts to various librarians. And all this for the sake of one property – the prediction of the READING relation for the reader, and, as a rule, for those objects (books) with which he had not previously entered into this relation!

REFERENCES

1. Polyakov, O.M. (2019), "Linguistic Data Model for Natural Languages and Artificial Intelligence. Part 2. Identification", *DISCOURSE*, vol. 5, no. 5, pp. 99–113. DOI: https://doi.org/10.32603/2412-8562-2019-5-5-99-113.

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

3. Polyakov, O.M. (2019), "Linguistic Data Model for Natural Languages and Artificial Intelligence. Part 3. Recognition", *DISCOURSE*, vol. 5, no. 6, pp. 132–143. DOI: https://doi.org/10.32603/2412-8562-2019-5-6-132-143.

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 33 scientific publications. Areas of expertise: linguistics, artificial intelligence, mathematics, database design theory, philosophy. E-mail: road.dust.spb@gmail.com

СПИСОК ЛИТЕРАТУРЫ

1. Polyakov O. M. Linguistic Data Model for Natural Languages and Artificial Intelligence. Part 2. Identification // DISCOURSE. 2019. Vol. 5, № 5. P. 99–113. DOI: 10.32603/2412-8562-2019-5-5-99-113.

2. Polyakov O. M. Linguistic Data Model for Natural Languages and Artificial Intelligence. Part 1. Categorization // DISCOURSE. 2019. Vol. 5, № 4. P. 102–114. DOI: 10.32603/2412-8562-2019-5-4-102-114.

3. Polyakov O. M. Linguistic Data Model for Natural Languages and Artificial Intelligence. Part 3. Recognition // DISCOURSE. 2019. Vol. 5, № 6. P. 132–143. DOI: https://doi.org/10.32603/2412-8562-2019-5-6-132-143.

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 33 научных публикаций. Сфера научных интересов: лингвистика, искусственных интеллект, математика, теория проектирования баз данных, философия. E-mail: road.dust.spb@gmail.com