Interpretability in Mathematics and XAI

Alexey Stukachev

(joint work with **Vitaliya Strepetova, Ulyana Zaitseva**,

Lyu Chen and **Tsao Tsylu**)

Sobolev Institute of Mathematics
Novosibirsk State University

MathAI 2025

Outline

We consider one of the central notions of eXplainable Artificial Intelligence, the notion of *interpretability*. This notion is actively and fruitfully used in mathematics. We discuss recent results on interpretability of object and processes in *mathematical and computational linguistics*. These two approaches to understanding and interpreting the meaning of a sentence or a text expressed in natural language provide very typical examples of differences between mathematical (theoretical) and computational (practical) models. Some examples of interpretations of models of one type in models of another type are presented.

Interpretation as a “semantic translation”

Types of interpretations discussed in this talk:

1. extension of existing mathematical model to formalise and explain some issues of natural language semantics
2. interpretation of one (theoretical) mathematical model of natural language semantics in another (more practical one)
3. adaptation of existing theoretical mathematical model to practical tasks of NLP

Remark. Effective (computable) interpretations can be formalised as, for example, Σ-reducibility (i.e., interpretation using Σ-formulas)

Outline

1. An example of formal mathematical interpretation of *Negative Concord*, typical for Russian and some other languages (but not for English).
2. An example of interpretation of formal mathematical model of *Montague semantics* for natural languages in practically oriented model of *DisCoCat*.
3. An example of interpretation (or adaptation) of formal model-theoretical notions of pair and union suitable for *datasets* used in computational linguistics.

1. NC: Skolem Functions vs Generalized Quantifiers

In mathematical linguistics, Negative Polarity Items (NPIs) and Negative Concord pose significant challenges in formal analysis of their semantics. We compare two approaches to the logical interpretation of NPIs and Negative Concord, based on generalized quantifiers and Skolem functions, respectively.

Natural languages feature expressions that appear exclusively in "negative" contexts, known as negative polarity items (NPIs), as well as positive polarity items (PPIs), which exist solely in "positive" contexts.

1. NC: Skolem Functions vs Generalized Quantifiers

Negative Concord is a phenomenon in which several formally negative units in a sentence express a single semantic negation.

NC is observed in many languages; e.g. Romance, Slavic, Greek, Hungarian, nonstandard English, West Flemish, Afrikaans, Lithuanian, Japanese

The question of how negative concord can undergo compositional semantic analysis has been explored in both syntactic and semantic literature. According to the principle of compositionality, standard English is characterized as a DN-language, for which (1a) is equivalent to (1b), while this is not the case for other languages, e.g., Russian.

(1a) Nobody didn't sleep.

(1b) Everybody slept.

The presence of two negations within a predicate logic formula renders it truth-conditionally equivalent to a positive formulation in (2), considering the logical law of double negation, which posits that two logical negations cancel each other.

1. NC: Skolem Functions vs Generalized Quantifiers

(2) $\neg\exists x[person(x) \wedge \neg sleep(x)] \Leftrightarrow \forall x[person(x) \rightarrow sleep(x)]$

Also, issues arise when considering languages with strict NC. The synonymous sentences in (3a) and (3b) can be represented by the logical form (4a). However, the logical form (4b) can be wrongly assigned to (3b), by compositions of its parts, because in Russian the combination of a negative pronoun with sentential negation does not yield a positive meaning:

(3a) Nobody slept

(3b) Nikto ne spal

(4a) $\lambda P.\neg\exists x[person(x) \wedge P(x)](\lambda v.sleep(v))$

(4b) $\lambda P.\neg\exists x[person(x) \wedge P(x)](\lambda v.\neg sleep(v))$

This conflict between the compositionality derived meaning and the actual interpretation of a sentence with NC illustrates the challenge that NC constructions pose for formal semantics.

1. NC: Skolem Functions vs Generalized Quantifiers

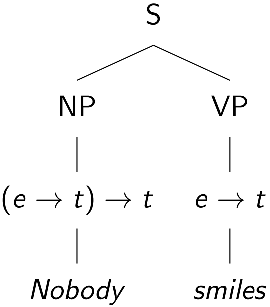
Montague in PTQ presents an approach to formal semantics of quantifiers in natural language. This analysis is based on Frege Compositionality Principle: "The meaning of a complex expression derives from the meanings of its constituent parts and their syntactical arrangement". He proposes to consider a nominal group as a collection of properties: the term *John* corresponds to the set of his properties, while *each person* refers to the set of qualities universally shared by people.

Barwise and Cooper deal with the integration of generalized quantifiers theory with linguistic phenomena. A distinctive feature of this theory is a formal representation of quantifier expressions.
$[\text{every student}]^e = \{P \subseteq D_e \mid [\text{student}] \subseteq P\}$
$[\text{most cats}]^e = \{P \subseteq D_e \mid |[\text{cat}] \cap P| > |[\text{cat}] \setminus P|\}$

1. NC: Skolem Functions vs Generalized Quantifiers

Consider the sentences “Nobody smiles” and “Nikto ne ulybaetsya” to identify the challenges with analiying expressions in sentences with negative concord. We construct a formal representation of the English sentence using lambda functions and a syntactic tree with the indication of expression types, taking into account that “nobody” in this analysis is a generalized quantifier. The similar analysis fo the second sentence is impossble!

$$\lambda P.\neg\exists x.P(x)(\lambda x.smile(x))$$



S

NP

VP

$(e \to t) \to t$

$e \to t$

Nobody

smiles

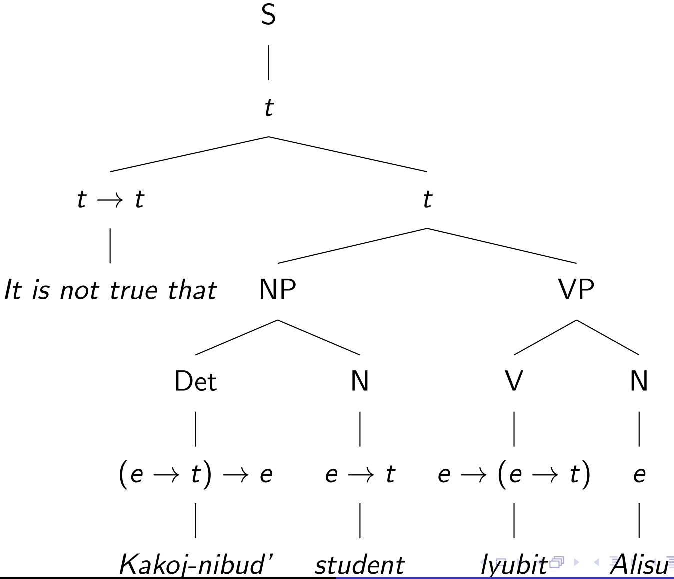
1. NC: Skolem Functions vs Generalized Quantifiers

Skolemization is a procedure by which existential quantifiers in a formula are eliminated. If the only quantifier in a formula is an existential quantifier, a Skolem constant is introduced in place of the variables bound by the existential quantifier: $\exists x P(x)$ is converted to $P(c)$. If the existential quantifier is preceded by the universality quantifiers, a new functional symbol f (Skolem function) is introduced: $\forall y \exists x (P(y, x))$ is transformed into $\forall y (P(y, f(y)))$.

Skolem functions cannot be directly applied to negating existential statements; however, they are necessary for understanding how to interpret existential statements in affirmative formulations.

1. NC: Skolem Functions vs Generalized Quantifiers

“Nikakoj student ne ljubit Alisu”



S

t

$t \to t$ — t

It is not true that — NP — VP

Det — N — V — N

$(e \to t) \to e$ — $e \to t$ — $e \to (e \to t)$ — e

Kakoj-nibud’ — *student* — *lyubit* — *Alisu*

1. NC: Skolem Functions vs Generalized Quantifiers

We indicate two main points of our analysis of generalized quantifiers and Skolem functions which show the latter are more suitable for the case of Negative Concord in Negative Polarity Items:

1) generalized quantifiers usually play a central role in the semantics of sentences (in particular, in monotonicity reasonings and entailments), while Skolem functions are technical in principle and allow other parts of a sentence (like verbs) to remain central;
2) functional types of Skolem functions $(e \to t) \to e$ are simpler than functional types of generalized quantifiers $(e \to t) \to ((e \to t) \to t)$ and have nothing on common with truth values and negation.

References

- Artem Burnistov and Alexey Stukachev, Generalized computable models and Montague semantics, *Studies in Computational Intelligence*, v. 1081 (2023), pp. 107-124.
- Artem Burnistov and Alexey Stukachev, Computable functionals of finite types in Montague semantics, *Siberian Electronic Mathematical Reports*, v. 21, N2 (2024), pp. 1460-1472.
- Ulyana Penzina and Alexey Stukachev, Skolem functions and generalized quantifiers for negative polarity items semantics, *Lecture Notes in Networks and Systems*, v. 1198 (2025), pp. 123-132.

References

- Montague, R.: The Proper Treatment of Quantification in Ordinary English, Approaches to Natural Language, . 221-242. Reprinted in Fml Philosophy: Selected Papers of Richard Montague, 1973, 247-270.
- Montague, R.: English as a forml language, Linguaggi nell nell Tecnica: pp 189-224. Reprinted in Foml Philosophy: Selected Papers of Richard Montague, 1974, 108-121.
- Montague, R.: Universal grammar, Theoria 36: 373-98. Reprinted in Formal Philosophy: Selected Papers of Richard Montague, 1974, 222-246.
- Coecke, B., Sadrzadeh, M., Clark, S.: Mathematical foundations for a compositional distributional model of meaning. arXiv preprint arXiv:1003.4394 (2010) https://arxiv.org/abs/1003.4394
- Toumi, A., de Felice, G.: Higher-Order DisCoCat (Peirce-Lambek-Montague semantics). arXiv preprint arXiv:2311.17813, (2023) https://arxiv.org/abs/2311.17813
- Tull, S., Lorenz, R., Clark, S., Khan, I., and Coecke, B.: Towards compositional interpretability for XAI (2024) https://arxiv.org/abs/2406.17583

2. Montague Semantics in DisCoCat

The *Compositionality Principle*, famously formulated by Frege, states that "the meaning of a complex expression is determined by the meanings of its parts and the rules used to combine them". This principle serves as the foundation for formal semantic frameworks, including Montague Semantics, which provides a rigorous, model-theoretic approach to natural language meaning using λ-calculus and intensional logic. Montague's work established a powerful framework for capturing the semantics of natural language through function application.

2. Montague Semantics in DisCoCat

At the same time, compositional models of meaning have seen significant development in the domain of distributional semantics, particularly with the DisCoCat (Categorical Distributional Compositional) model. This approach combines distributional semantics, which represents word meanings via vector spaces, and compositionality principle via category theory. DisCoCat models have gained popularity in recent years, not only within computational linguistics but also in fields such as quantum computing and explainable artificial intelligence (XAI).

2. Montague Semantics in DisCoCat

Emerging from the philosophy of language, Montague Semantics excels in theoretical precision but lacks practical applicability in data-driven contexts, its intensional logic and functional structures do not naturally lend themselves to computational implementation at scale. In contrast, DisCoCat model provides a concrete framework for computational applications, benefiting from the advances in machine learning and vector-based representations. However, in its standard form, it lacks the flexibility needed to model complex linguistic phenomena, such as higher-order modifications, quantification, and intensionality.

2. Montague Semantics in DisCoCat

Recent research has explored ways to connect these two frameworks, with Higher-Order DisCoCat (HO-DisCoCat) serving as a notable example. By incorporating elements from Montague Semantics, such as λ-calculus, HO-DisCoCat extends the expressive power of distributional semantics, allowing for more refined compositional mechanisms. This development highlights the potential for a deeper integration between formal logical approaches and distributional models of meaning.

We present an effective interpretation of a fragment of Montague Semantics in DisCoCat (in particular HO-DisCoCat). This opens up new prospects for analysing, understanding, and refining existing approaches.

2. Montague Semantics in DisCoCat

Higher-Order DisCoCat was introduced by A.Toumi and G.de Felice, with the meaning of a word interpreted by a diagram-valued higher-order function. That allows to treat higher-order and non-linear issues in natural language semantics: adverbs, prepositions, negation and quantifiers. This approach connects HO-DisCoCat method with Montague semantics, since it can be seen as a variant of λ-calculus where the primitives act on string diagrams.Three main directions to compose diagrams:

1. *left-to-right* using the composition $\circ_{xyz} : (x, y) \times (y, z) \to (x, z)$ which connects the output of one diagram to the input of another (*sequential*)
2. *top-to-bottom* using the tensor $\otimes_{xyzw} : (x, y) \times (z, w) \to (xz, yw)$ which concatenates two diagrams side by side (*parallel*)
3. *inside-out* using the evaluation $(x, y) \times [(x, y) \to (z, w)] \to (z, w)$ which substitutes one diagram for the free variable in another (inner)

2. Montague Semantics in DisCoCat

The classical DisCoCat model utilises only the first two directions, limiting its ability to express more complex semantic phenomena. In HO-DisCoCat, the third direction, the ability for diagrams to contain other diagrams as embedded structures, enables the modelling of higher-order functions, such as adverbs and other modifiers that influence sentence structure.

It is possible to represent second-order constructions, such as adverbs, as boxes with k holes that can contain diagrams of a certain shape. We can define *monoidal signature with holes*:

$$\Sigma = (\Sigma_0, \Sigma_1, dom, cod, holes)$$

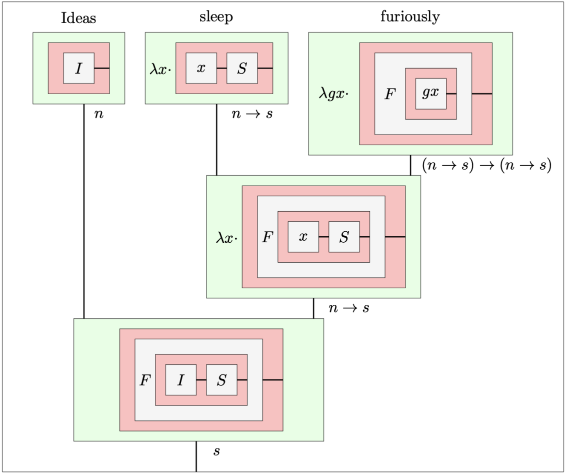
as a monoidal signature together with a function $holes : \Sigma_1 \to (T \times T)^*$ which assigns to each box $(f : x \to y) \in \Sigma$ a list of pairs of types:

$$holes(f) = ((z_1, \omega_1), \ldots, (z_k, \omega_k)) \in (T \times T)^*$$

for the domain and codomain of the diagrams that can fit in its $k \in \mathbb{N}$ holes.

2. Montague Semantics in DisCoCat

For example, in the sentence "Ideas sleep furiously" the semantics of adverbs like "furiously" can now de defined as boxes with one hole for a verb, prepositions like "with" as boxes with two holes for nouns, etc.



References

- Artem Burnistov and Alexey Stukachev, Generalized computable models and Montague semantics, *Studies in Computational Intelligence*, v. 1081 (2023), pp. 107-124.
- Artem Burnistov and Alexey Stukachev, Computable functionals of finite types in Montague semantics, *Siberian Electronic Mathematical Reports*, v. 21, N2 (2024), pp. 1460-1472.
- Ulyana Penzina and Alexey Stukachev, Skolem functions and generalized quantifiers for negative polarity items semantics, *Lecture Notes in Networks and Systems*, v. 1198 (2025), pp. 123-132.

References

- Montague, R.: The Proper Treatment of Quantification in Ordinary English, Approaches to Natural Language, . 221-242. Reprinted in Fml Philosophy: Selected Papers of Richard Montague, 1973, 247-270.
- Montague, R.: English as a forml language, Linguaggi nell nell Tecnica: pp 189-224. Reprinted in Foml Philosophy: Selected Papers of Richard Montague, 1974, 108-121.
- Montague, R.: Universal grammar, Theoria 36: 373-98. Reprinted in Formal Philosophy: Selected Papers of Richard Montague, 1974, 222-246.
- Coecke, B., Sadrzadeh, M., Clark, S.: Mathematical foundations for a compositional distributional model of meaning. arXiv preprint arXiv:1003.4394 (2010) https://arxiv.org/abs/1003.4394
- Toumi, A., de Felice, G.: Higher-Order DisCoCat (Peirce-Lambek-Montague semantics). arXiv preprint arXiv:2311.17813, (2023) https://arxiv.org/abs/2311.17813
- Tull, S., Lorenz, R., Clark, S., Khan, I., and Coecke, B.: Towards compositional interpretability for XAI (2024) https://arxiv.org/abs/2406.17583

3. Adaptation of Model-Theoretical Notions for NLP Tasks

As the importance of machine learning continues to increase in various fields, the quality and consistency of datasets have become key issues, especially when computing power is limited. Labelling datasets has been proven to be an effective means to improve training efficiency and accuracy, providing support for improving the accuracy and reliability of data analysis.

We recall the basic components of datasets, including vocabulary, part-of-speech tags, relation tags, and their labelling functions. By introducing contradictory sets, reflexive sets, sequential sets, and general sets, we ensure the integrity and consistency of datasets.

3. Adaptation of Model-Theoretical Notions for NLP Tasks

For merging datasets, we describe the construction methods for unions and intersections, and define new part-of-speech and relation labelling functions to process elements from different datasets. In particular, our research provides a theoretical basis for the effective merging of datasets.

A dataset is an organized collection of (linguistic) data. In supervised learning, the labelled target outputs of data instances are key to training models. The labelling content depends on the specific task type. In the field of natural language processing, some datasets label the parts of speech of words, and some datasets label the semantic relations between words. Complex models may have multiple labels simultaneously. Assuming a dataset has parts of speech and semantic relations as labels, here is the mathematical definition of this dataset.

3. Adaptation of Model-Theoretical Notions for NLP Tasks

Let $\mathcal{D}$ be a dataset, then:
$\mathcal{D} = (V, P, R, L_p, L_r, S)$
where:

- V is a vocabulary, $V = \{w_1, w_2, \ldots, w_n\}$, w_i represents a word.
- P is a set of part-of-speech tags, $P = \{p_1, p_2, \ldots, p_m\}$, p_i represents a part of speech.
- R is a set of relation labels, $R = \{r_1, r_2, \ldots, r_k\}$, r_i represents a binary relation.
- $L_p : V \to P$ is a part-of-speech labelling function.
- $L_r : V \times V \to \mathcal{P}(R)$ is a relation annotation function.
- S is a set of sentences, $S = \{s_1, s_2, \ldots, s_t\}$.

3. Adaptation of Model-Theoretical Notions for NLP Tasks

For any sentence s, we can present it as

$$s = (W, E),$$

where

- $W = (w_1, w_2, \dots, w_n)$ is a sequence of words, $w_i \in V$.
- $E = \{(i, j, r) \mid i, j \in [1, n], r \in \mathcal{P}(R)\}$ is a set of relation edges.

and the following two constraints need to be met:

1. Labelling constraints for part of speech: for all words $w \in V : L_p(w) \in \mathcal{P}(P)$
   That is, words may not have part-of-speech labels.
2. Labelling constraints for relations: for all pairs $(w_1, w_2) \in V \times V : L_r(w_1, w_2) \in \mathcal{P}(R)$
   That is, there may be no connection between these two words.

3. Adaptation of Model-Theoretical Notions for NLP Tasks

With the set of relation labels, we define four more sets, namely, the contradictory set A, the reflexive set B, the sequential set C and the general set G. This is necessary to ensure the correctness of the dataset. These four sets need to be manually generated based on the relations in the relation label set. We then need to check the dataset against the generated set and make corrections if any inconsistencies are found. The relations contained in the following sets are all from the relation label set.

Contradictory set A: contains all mutually exclusive pairs of relation labels.

Here are some examples:

- (parent, child): x cannot be both the parent of y and the child of y.
- (greater than, less than): x cannot be both greater than and less than y.

3. Adaptation of Model-Theoretical Notions for NLP Tasks

For set A, we calculate the number of occurrences of each contradictory relation separately and use it as the deletion level of this relation.
For example, for set of relation labels R, we generate a contradiction set A :

$$A = \{\{r_1, r_2\}, \{r_1, r_3\}, \{r_1, r_5\}, \{r_2, r_4\}, \{r_3, r_5\}\}$$

The deletion level of r_1 is 3, r_2 is 2, r_3 is 2, r_4 is 1, and r_5 is 1 .

References

1. Montague, R.: The Proper Treatment of Quantification in Ordinary English, Approaches to Natural Language, . 221-242. Reprinted in Fml Philosophy: Selected Papers of Richard Montague, 1973, 247-270.
2. Montague, R.: English as a forml language, Linguaggi nell nell Tecnica: pp 189-224. Reprinted in Foml Philosophy: Selected Papers of Richard Montague, 1974, 108-121.
3. Montague, R.: Universal grammar, Theoria 36: 373-98. Reprinted in Formal Philosophy: Selected Papers of Richard Montague, 1974, 222-246.
4. Coecke, B., Sadrzadeh, M., Clark, S.: Mathematical foundations for a compositional distributional model of meaning. arXiv preprint arXiv:1003.4394 (2010) https://arxiv.org/abs/1003.4394
5. Toumi, A., de Felice, G.: Higher-Order DisCoCat (Peirce-Lambek-Montague semantics). arXiv preprint arXiv:2311.17813, (2023) https://arxiv.org/abs/2311.17813
6. Tull, S., Lorenz, R., Clark, S., Khan, I., and Coecke, B.: Towards compositional interpretability for XAI (2024) https://arxiv.org/abs/2406.17583

References

- J. Barwise, Admissible Sets and Structures, Springer Verlag, 1975
- Yu.L. Ershov, Definability and Computability, Plenum, 1996
- Alexey Stukachev, Effective Model Theory: an approach via Σ-Definability, Lecture Notes in Logic, v. 41 (2013), pp. 164-197
- Artem Burnistov and Alexey Stukachev, Generalized computable models and Montague semantics, *Studies in Computational Intelligence*, v. 1081 (2023), pp. 107-124.
- Artem Burnistov and Alexey Stukachev, Computable functionals of finite types in Montague semantics, *Siberian Electronic Mathematical Reports*, v. 21, N2 (2024), pp. 1460-1472.
- Ulyana Penzina and Alexey Stukachev, Skolem functions and generalized quantifiers for negative polarity items semantics, *Lecture Notes in Networks and Systems*, v. 1198 (2025), pp. 123-132.

Thank You!