Igor Boguslavsky

LEXICAL RESOURCES FOR A SEMANTIC PARSER OF RUSSIAN

Argument Structure of Ordinal Adjectives

Abstract The SemETAP semantic model is a part of a more general ETAP linguistic processor aiming at analyzing and generating NL texts. The task of SemETAP is building two kinds of semantic structures – Basic SemS, which capture the core meaning of the sentence, and Enhanced SemS, which contain diverse inferences drawn from Basic SemS. SemETAP is supported by two main lexical resources – a combinatory dictionary of Russian and an ontology. One important requirement for the SemS is that it should explicitly represent all semantic arguments of the predicates of the sentence, expressed by all kinds of words – verbs, nouns, adjectives or adverbs. We discuss the argument structure of ordinal adjectives (*first, second,..., last, next*), which has been largely neglected in the literature on valency and arguments. Several semantic slots are introduced for ordinal adjectives: hasObject, hasObject2, belongsTo, hasNumber, hasStartingPoint, hasTerminalPoint, orderedBy. Our analysis reveals interesting features in the behavior of the arguments of ordinal adjectives.

Keywords semantic structure; combinatorial dictionary; ontology; argument structure; ordinal adjectives

1. Semantic Parser SemETAP

This work is part of the development of lexical resources elaborated for the SemETAP functional semantic model (Boguslavsky, Frolova et al., 2018; Boguslavsky, Frolova et al. 2019; Boguslavsky, Dikonov et al., 2020). Its immediate task is building semantic representations of Russian texts. Semantic structures (SemS) constructed by SemETAP aim to represent the meaning of the sentence in an explicit and transparent way, while abstracting away from the lexico-syntactic variation of the NL. The model has several salient features.

- SemETAP aims to extract a large number of inferences from a text. We assume that the more inferences we can extract from a text, the more complete and deeper understanding of the text we achieve.
- The model distinguishes between two levels of semantic representation of a sentence: The Basic Semantic Structure, which captures the core meaning of the sentence, and the Enhanced Semantic Structure, which enriches the Basic Structure with a large number of implications. The extension of the Basic Structure is carried out through two main sources: decomposition of concept meanings and common-sense axioms.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

- There are two types of implications: strict implicatures and plausible expectations. The latter are particularly important for understanding coherent text and recovering implicit elements of meaning.
- The model is knowledge-based, which means that it relies on explicit knowledge provided by an expert. There are two main lexical resources – a combinatory dictionary of Russian and an ontology, which can be thought of as a semantic dictionary. Both types of semantic structures – the Basic and the Enhanced ones – are built of the elements of the ontology.
- The model functions as a component of a larger system of analysis and generation of texts known as ETAP, which has also a morphological and a syntactic module. In this way, a complete sentence processing cycle is carried out, from the text in orthographic notation to the Enhanced Semantic Structure.
- Basic semantic structures built according to this model constitute a novel semantic corpus (SemOntoCor) that contains more than 2000 sentences at the time of writing and keeps growing (Boguslavsky, Dikonov et al., 2023).

Here is an example to illustrate what SemETAP can do. Sentence (1) is transformed into the Basic SemS shown in Figure 1. (For clarity, green boxes are used to show the correspondence between the words of the sentence and the fragments of the structure).

(1) *Роналду не спас матч.*
'Ronaldo did not save the match'

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

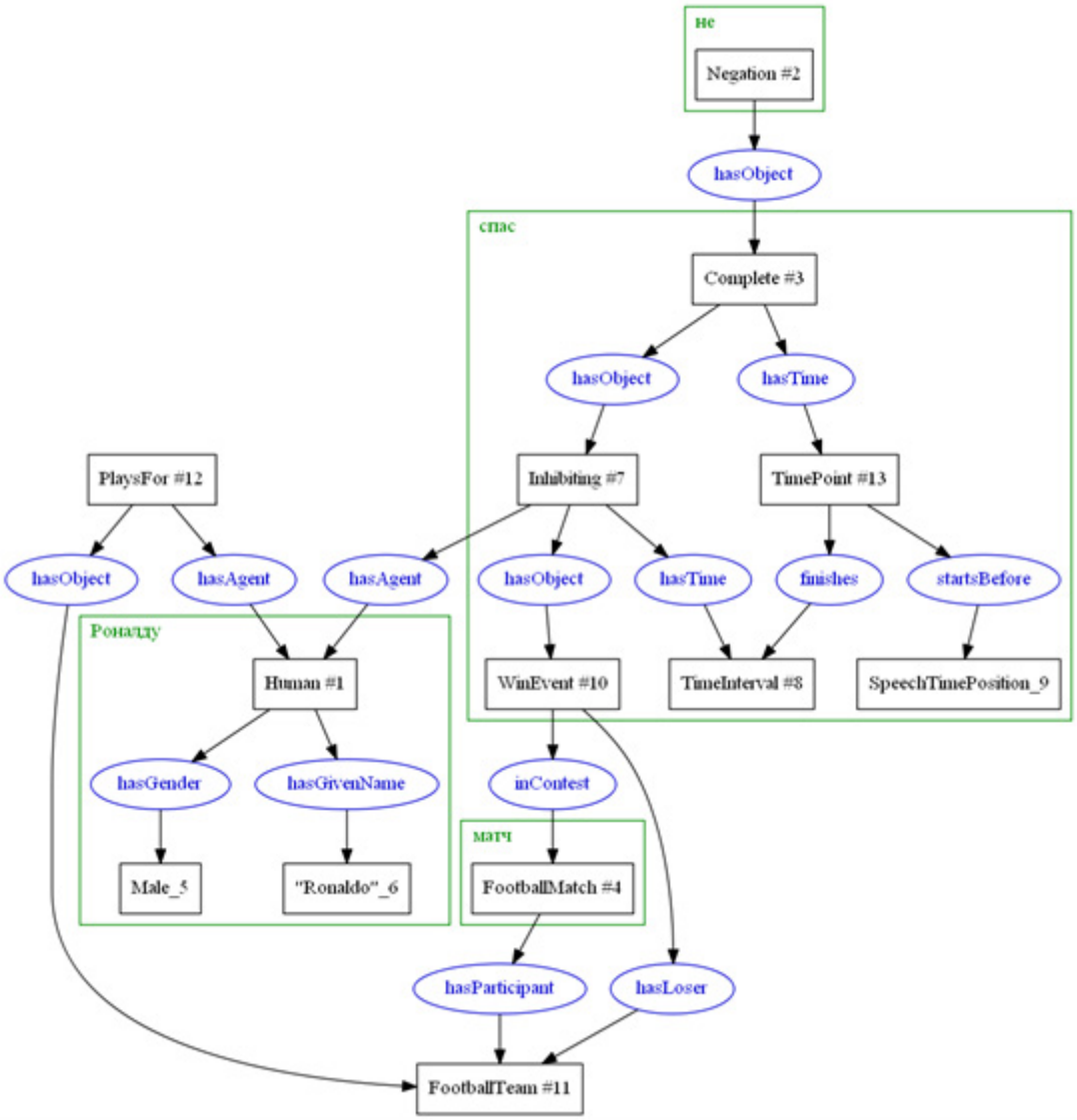

Fig. 1: The Basic semantic structure of the sentence *Роналду не спас матч* 'Ronaldo did not save the match'.

This structure can be "read" as follows: "human #1, whose name is Ronaldo, who is a male and plays for FootballTeam #11, did not inhibit defeat of FootballTeam #11 in a football match in which it participated. All this happened in the past (= before the time of speech)".

To demonstrate the SemETAP ability to draw inferences, it is convenient to make use of the question-answering mode of the system. In Figure 2, one can see the result of processing sentence (1).

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

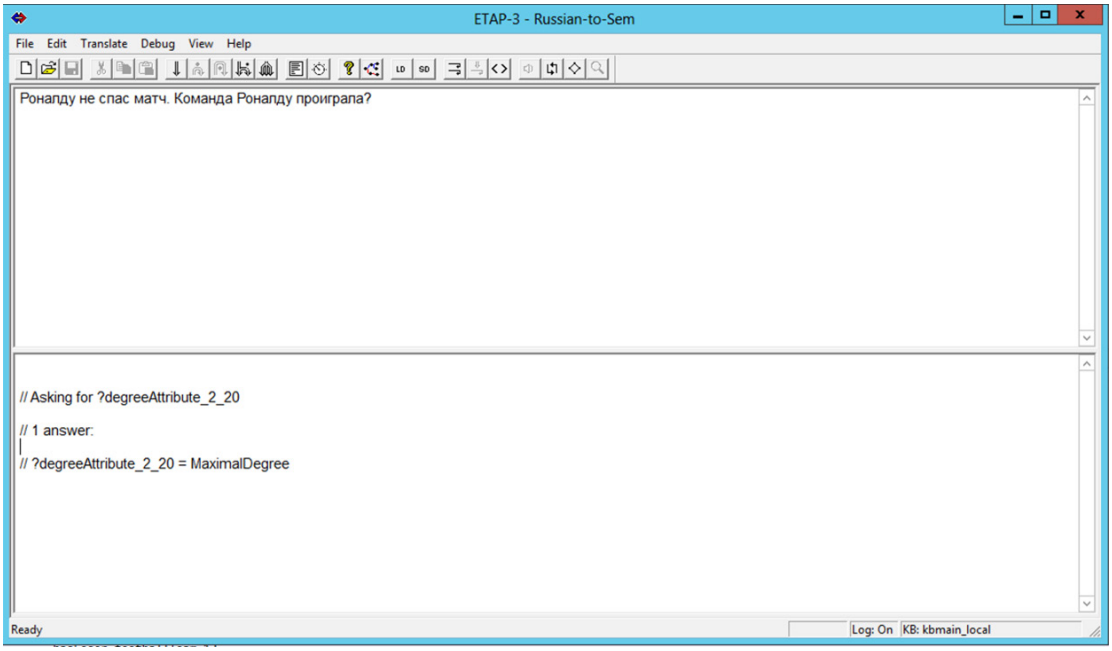

Fig. 2: Sentence (1) and the question *Команда Роналду проиграла?* 'Did Ronaldo's team lose the match?'

In the upper window of Figure 2 is the text (*Ronaldo did not save the match*) and the diagnostic question (*Did Ronaldo's team lose the match?*). The lower window contains the answer returned by the system. Let us explain this answer. For SemETAP, finding if a proposition is true amounts to discovering the value of the epistemic modality of this proposition. The meaning of the diagnostic question is: What is the value of the attribute ?degreeAttribute_2_20, which is the value of the epistemic modality of the statement "the team for which Ronaldo was playing lost the match". In plain words, is it true that Ronaldo's team lost the match? The answer, which can be seen in the lower window, reads that the value of this modality is maximal. This means that the question is answered in the affirmative.

2. Ordinal Adjectives and Their Argument Structure

As mentioned above, semantic structures (SemS) constructed by the analyzer aim to represent the meaning of the sentence in an explicit and transparent way, while abstracting away from the lexico-syntactic variation of the NL. One important requirement for the SemS is that it should explicitly represent all the semantic arguments of the predicates of the sentence, expressed by all kinds of words – verbs, nouns, adjectives or adverbs (for more on the typology of different types of arguments see Boguslavsky, 2014). SemS is not constructed from NL words, but from semantic elements (concepts), which are connected by special semantic relations. Accordingly, the analyzer has two lexical resources – a semantic dictionary (ontology) and a combinatorial dictionary of Russian. All lexical information used by the analyzer to construct the SemSs is distributed between these two resources. The ontology contains an inventory of concepts and describes their properties - in particular, their argument structure. The combinatorial dictionary describes the properties of words and the relationship between these words and the ontology.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

The lexical resources of the semantic analyzer should make it possible to solve the following tasks:

a) Introduce the concepts (or combinations of concepts) necessary and sufficient for representing NL words and define their argument structure,

b) Establish the correspondence between Russian words and ontological concepts expressing them in different contexts,

c) Specify the ways in which each argument of these concepts is realized in the texts.

The first task is solved in the ontology, while the second and the third ones, in the combinatorial dictionary of Russian.

The subject of this paper are Russian ordinal adjectives and some related words (OAdj), which denote the position of an object in an ordered sequence of homogeneous objects, such as *first, second, third, ... last, penultimate, next, previous.* Their meaning is transparent enough. For example, adjective *пятый* 'fifth' in *пятая книга* 'the fifth book' means that the book under discussion occupies the fifth position in some ordered sequence of books. Surprisingly, the argument structure of these adjectives has not attracted due attention of either theoretical linguists or lexicographers of different languages and has not been described in the available literature and lexical resources with the necessary completeness (Geist, 2010; Graschchenkov, 2008; Graschchenkov & Kobozeva, 2017; Gutiérrez et al., 2005; Ikeya, 1995; Johnson & Fillmore, 2000; Meltzer-Asscher, 2011; Proceedings, 2022). In this paper we will try to fill this gap. It is obvious that in order to fully specify the position of an object in an ordered sequence, in a general case it is necessary to take into account several different aspects of the situation. For example, in the phrase *третий дом от аптеки* 'the third house after the pharmacy', the phrase 'after the pharmacy' signals where the sequence of numbered elements begins. The first element of this sequence is the house immediately following the pharmacy. Thus, several arguments (semantic slots) should be postulated for the concept of numbering objects and for the corresponding OAdj.

Below, we will enumerate the concepts proposed for OAdj and describe their argument structure. The ways in which arguments are expressed will be evident from the examples given. After that, we will comment on some non-obvious properties of OAdj.

The concepts used to represent OAdj are HavingOrderedPosition, Last, Penultimate, Next and Previous. All of them make part of the ontological class OrderingRelation. Below, we list their semantic arguments, indicating for each argument the semantic relation by which it attaches to the corresponding concept, a brief explanation, and examples.

Concept HavingOrderedPosition.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

Arguments:

- hasObject – An entity having a number: *третий том* 'third volume', *пришел в третий раз* 'came for the third time'.
- belongsTo – A sequence (or a domain) within which the object is thought of: *первая скамейка на бульваре* 'first bench on the boulevard', *седьмой в очереди* 'seventh in the queue', *третий альбом на верхней полке* 'third album on the top shelf', *второй дом по левой стороне улицы* 'second house on the left side of the street', *первый успех за 10 лет* 'first success in 10 years', *первая победа в турнире / из 10 матчей* 'first win in a tournament/out of ten matches'; *второй раз за этот день* 'second time this day'.
- hasNumber – The number of the object in this sequence.
- orderedBy – Criterion for ordering objects in a sequence: *третий по росту* 'third in height'. This criterion may not be explicit. By default, events are ordered by time: *вторая победа* 'second victory' took place earlier than *третья победа* 'third victory'. Physical objects are often ordered by their location relative to the observer: *третий дом по левой стороне* 'the third house on the left hand side' is closer to the observer than *четвертый дом* 'the fourth house'.
- hasStartingPoint - Position from which starts the counting: *третий дом от аптеки (от конца)* 'third house from the pharmacy (from the end)', *третья строка снизу* 'third line from the bottom', *первый праздник после Нового года* 'first holiday after the New Year').
- hasTerminalPoint – Direction of counting: *пятая остановка в сторону центра* 'fifth stop towards downtown'.

Here are some examples of SemS, which show how OAdj express their arguments in the text and how they are presented in semantic structures.

(2) *Третий дом от аптеки в сторону центра* 'the third house from the pharmacy towards downtown'

House
    isObjectOf HavingOrderedPosition
        hasNumber 3
        hasStartingPoint DrugStore
        hasTerminalPoint Downtown

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

For illustration purposes, this structure is visualized below by means of a graph:

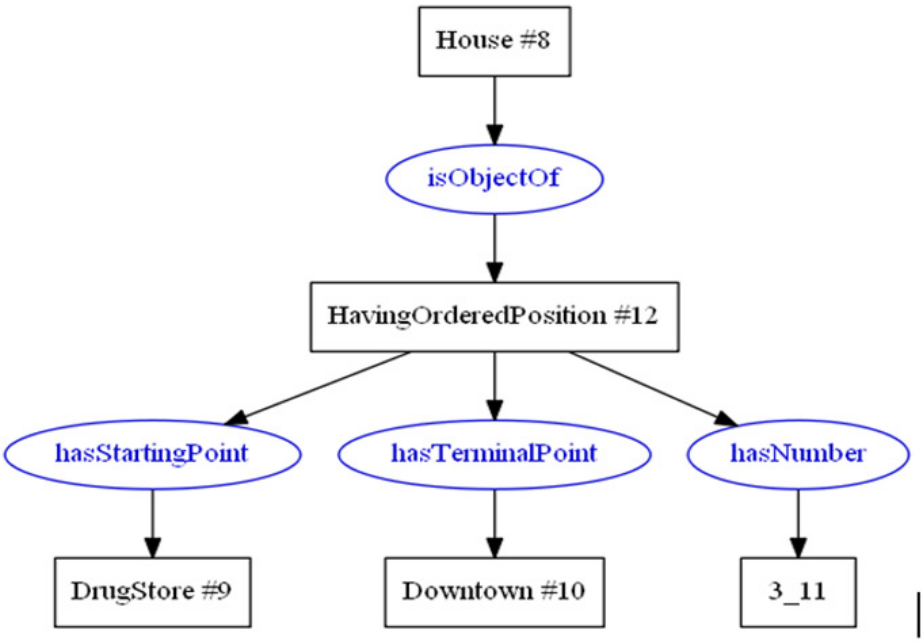

Fig. 3: The Basic SemS of the phrase *Третий дом от аптеки в сторону центра* 'the third house from the pharmacy towards downtown'

(3) *Домик стоял четвертым слева на последней от штаба улице* 'The cottage stood fourth on the left on the last street from the headquarters'.

```
HavingOrderedPosition
        hasObject House
                isObjectOf Location
                        hasObject2 Street
                                isObjectOf Last
                                        hasStartingPoint Headquarters
        hasNumber 4
        hasStartingPoint Left
```

(4) *Первый ученик в классе по росту 'the first student in the class in height (=the tallest)'*

```
Student
        isObjectOf HavingOrderedPosition
                hasNumber 1
                belongsTo Class
                orderedBy HavingHeight
```

The Last and Penultimate concepts have the same arguments as HavingOrderedPosition except that they do not have the argument hasNumber. On the other hand, Next has a specific argument position ('next after' - hasObject2) absent in HavingOrderedPosition:

(5) *Экзамен, следующий в этой сессии по важности после математики* 'the next most important exam in this session, after maths'.

Exam
    isObjectOf Next
        hasObject2 Mathematics
        orderedBy Importance
        belongsTo Session

3. Some Linguistic Peculiarities of OAdj Arguments

3.1 Starting Point of Numbering

The starting point argument shows where one should start counting from, but it may not be as simple as it seems. Let us compare sentences (6) and (7), which seem very similar:

(6) *Моя школа – второй дом после аптеки.*

'my school is the second house after the pharmacy'

(7) *Вице-президент – второе лицо после президента.*

'the Vice President is the second official person after the President'

In both cases, an object is specified which serves as a reference point for numbering (*the pharmacy, the President*), but there is an important difference. In (6), there are three houses before our eyes: the pharmacy, my school, and the house between the two. By contrast, in (7) there is no third person between the President and the Vice President. The former is immediately followed by the latter.

The difference in the interpretation of (6) and (7) is that in (6) the pharmacy is not thought of as a member of the numbered sequence. It only marks the point **after which the sequence begins**. Therefore, the house next to the pharmacy is numbered 1 in the sequence, and my school gets number 2. In (7), the President is **the element that starts the sequence**. Therefore, he gets number 1 and the next element (the Vice President) gets number 2.

Let us formulate the difference between (6) and (7) in more general terms. Let element A be *n-th* after X (in our examples (6)-(7) *n=2*). Then the meaning of (6) can be represented as follows. Element A (*my school*) has number *n* (*2*) in the sequence that begins after X (*the pharmacy*). Between X and A there are *n-1* (*1*) unnamed elements of the sequence.

The meaning of (7) looks different. Element A (*Vice President*) has also number *n* (*2*) in a sequence, but X (*President*) is what constitutes the first *n-1* (*1*) elements of the sequence.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

The next difference between the above two variants of the reference point in (6) and (7) concerns the structure of the X group itself, that fills the hasStartingPoint slot. First of all, let us note that interpretation (7) requires a complete enumeration of all the first *n-1* elements of the sequence. (8) is an example where n=3, and therefore X has to enumerate the first 2 elements of the sequence:

(8) *Спикер Палаты представителей Конгресса – третье лицо после президента и вице-президента.*

'the Speaker of the House of Representatives of the Congress is the third person after the President and the Vice President'

In the next example n=4:

(9) *Россия является четвертым после Швеции, США и Дании государством по объему накопленных инвестиций* (Diplomatičeskij Vestnik, 2004).

'Russia is the fourth country after Sweden, the USA and Denmark in terms of accumulated investment'

Let us also note that the elements of the coordinated phrase X cannot be switched. Sentence (9) not only tells us that Russia ranks 4th in terms of accumulated investment. It also tells us that Sweden ranks first, the United States second, and Denmark third. The coordinated elements should follow exactly the order in which the countries are ordered in terms of investment volume. Changing the order of the coordinated elements does not make the sentence grammatically incorrect, but it does make it false.

As far as sentences of the type (6) are concerned, they do not accept coordinated strings as hasStartingPoint arguments. Sentence (10) is ungrammatical (unless the pharmacy and the department store are situated in the same building).

(10) **Моя школа – второй дом после аптеки и универмага.*

'my school is the second house after the pharmacy and the department store'

3.2 Numbered Sequence Includes the Main Predicate

As can be seen from the examples given, in a sentence OAdj arguments are usually expressed by groups syntactically subordinate to OAdj (with the exception of the numbered object itself, which is expressed by the noun being defined): *первый в классе* 'first in class', *второй по росту* 'second in height', *пятая строка снизу* 'fifth line from the bottom'. Note, however, a case where the argument denoting the sequence itself (introduced by the belongsTo relation) includes the main predicate of the clause.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

As a rule, OAdj together with their arguments are interpreted within the NP to which they belong. For example, in the sentence *Я купил билет в пятый вагон* 'I bought a ticket to the fifth carriage', the number of the carriage is defined without any connection with the verb *bought*. Meanwhile, there are sentences in which it is impossible to establish the sequence in which a given object is thought of without resorting to the main predicate of the clause. For example, the sentence (11)

(11) *Алькарас выигрывает уже третий турнир*

'Alcaras wins (his) third tournament'

is natural to understand in the sense that Alcaras has already won two tournaments, rather than in the sense that in a series of somehow ordered tournaments he wins tournament number 3. In other words, the sequence in which the tournament in question has number 3 consists of the tournaments won by Alcaras.

4. Conclusion

The task of the semantic analyzer is to represent the meaning of a text in an explicit and transparent way. Identifying the arguments of all predicates of a text is the main subtask of this complex task, since predicate-argument relations are the "semantic glue" that unites individual semantic elements into a coherent semantic structure of a sentence. We consider the predicate-argument structure of a sentence somewhat more broadly than is often done. To the class of argument-containing words we include many adjectives and adverbs that have not been sufficiently studied in lexicography from this point of view. A good example is the compact class of ordinal adjectives (*first, second, third,... last, next*). We show that in order to characterize the place of an object in an ordered sequence of similar objects it is necessary to take into account several parameters of the situation, such as the sequence within which the object is thought of (*the first bench on the boulevard*), the criterion for ordering objects in the sequence (*third in height*), the position from which the counting starts (*third line from the bottom*), and the direction of the counting (*fifth stop towards the city center*). As a rule, these parameters are expressed in the text by means of phrases syntactically connected with ordinal adjectives and meet all the requirements imposed on semantic valences (Mel'čuk, 2015). Some of these valences have interesting properties from the linguistic point of view (see Section 3).

References

Boguslavsky, I., Frolova, T., Iomdin, L., Lazursky, A., Rygaev, I., & Timoshenko, S. (2018). Semantic analysis with inference: high spots of the football match. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue", 17*(24), 124–142.

Boguslavsky, I. M., Frolova T. I., Iomdin L. L., Lazursky A. V., Rygaev, I. P., & Timoshenko, S. P. (2019). Knowledge-based approach to Winograd Schema Challenge. In *Computational linguistics and intellectual technologies. Papers from the Annual International Conference "Dialogue", 18*(25), 86–103.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

Boguslavsky, I. M., Dikonov, V. G., Frolova, T. I., Iomdin, L. L., Lazursky, A. V., Rygaev, I. P., & Timoshenko, S. P. (2020). Full-fledged semantic analysis as a tool for resolving Triangle-COPA social scenarios. In *Computational linguistics and intellectual technologies. Papers from the Annual International Conference "Dialogue", 19*(26), 106–118.

Boguslavsky, I. M., Dikonov, V. G., Frolova, T. I., Iomdin, L. L., Lazursky, A. V., Rygaev, I. P., & Timoshenko, S. P. (2023). Constructing a Semantic Corpus for Russian: SemOntoCor. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2023"*, *21*(28), 1–14.

Boguslavsky, I. (2014). Argument structure of adverbial derivatives in Russian. In J. Tsujii, & J. Hajic, *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* (pp. 1071–1080). Dublin, Ireland, August 23–29, 2014.

Geist, L. (2010). The argument structure of predicate adjectives in Russian. *Russian Linguistics, 34,* 239–260. https://doi.org/10.1007/s11185-010-9064-5

Grashchenkov, P. V. (2008). *The Argument Structure of Russian Adjectives (ms.).* Institute of Oriental Studies, Moscow, 37 p.

Grashchenkov, P. V., & Kobozeva I. M. (2017). Semantic classes and government of adjectives. In Selegei et al. (Eds.), *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue"*, *16*(23), 134–149.

Gutiérrez, E., Javier, L., & Fábregas Alfaro, A. (2005). Cuestiones de orden: semántica y sintaxis de los adjetivos ordinales. *ELUA. Estudios de Lingüística*, *19*, 85–105.

Ikeya, A. (1995). Predicate-Argument Structure of English Adjectives. In B. K. T'sou, & T. B. Y. Lai, *Proceedings of the 10th Pacific Asia Conference on Language, Information and Computation*, (pp. 149–156), City University of Hong Kong, Hong Kong. City University of Hong Kong.

Johnson C., & Fillmore Ch. (2000). The FrameNet tagset for frame-semantic and syntactic coding of predicate-argument structure. In *NAACL 2000: Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, April 2000, 56–62.

Mel'čuk, I. (2015). *Semantics. From Meaning to Text. Vol. 3.* John Benjamins Publishing Company, Amsterdam/Philadelphia.

Meltzer-Asscher A. (2011). *Adjectives and Argument Structure.* [Doctoral dissertation, Tel Aviv University].

Proceedings of the Workshop on Aspect and Argument Structure of Adjectives/Adverbs and Participles/Prepositions. (2022). University of Greenwich. https://docs.gre.ac.uk/__data/assets/pdf_file/0017/290042/waasap-10th-anniversary-conference-final.pdf

Acknowledgements

This work was supported by the RSF grant No. 24-18-00988, which is gratefully acknowledged.

Contact information

Igor Boguslavsky
Institute for Information Transmission Problems / Russian Academy of Sciences; Universidad Politécnica de Madrid.
Igor.M.Boguslavsky@gmail.com.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.