Dictionary as an Instrument of Linguistic Research

Valentina Apresjan, Nikolai Mikulin
National Research University "Higher School of Economics", School of Linguistics
Vinogradov Russian Language Institute
e-mail: vapresyan@hse.ru, nickmikulin@gmail.com

Abstract

While linguistic corpora have long been established as an efficient instrument of linguistic research, dictionaries are underrated in this capacity. Their utilitarian function is mostly defined by their usefulness to a general user. Accordingly, electronic dictionaries are primarily designed to satisfy the needs of a language learner, and not of a linguist. But dictionaries of an active type, with their wealth of structured multi-level information about words, could feed linguistic theory in a variety of ways. Active dictionary of the Russian language (ADR) is an advanced active dictionary aimed at recording all linguistic information necessary to use a word correctly. ADR includes information on the morphological, stylistic, prosodic, syntactic, and combinatorial properties of lexical items, as well as on their major semantic relations (synonyms, antonyms, converse synonyms, derivatives). The current project involves turning ADR into an electronic database with a variety of search functions, that would both advance linguistic theory and serve the needs of a general user. The paper is structured as follows: 1) General principles of ADR; 2) ADR Online as an electronic database; 3) Sample uses of ADR Online as an instrument of linguistic research.

Keywords: active dictionary; lexicographic type; semantic class; database; integrated description of language; negative polarity

1 General Principles of ADR

ADR is meant to function both as a scholarly description of language and a lexicographic reference book of an active type (Apresjan 2014). Unlike passive dictionaries that are mostly geared towards users who need to interpret or decode texts, active dictionaries are orientated towards encoding. This explains their quantitative and qualitative differences: while passive dictionaries comprise large word-lists with minimum information about each word, sufficient only for understanding it in a given context, active dictionaries feature significantly smaller lexical coverage but contain comprehensive information about each lexical item sufficient for using this item in a given context. As mentioned above, ADR's intended lexical coverage is about 12,000 lexical items.

1.1 Theoretical Basis of ADR

Theoretically, ADR is based on the principles of the integrated description of language and systematic lexicography (Apresjan 2000), in particular, on the principle of semantic motivation underlying the properties of linguistic items. The presence of a correlation between semantics and other properties of linguistic items means that linguistic items form semantic classes that share not only semantic closeness, but also possess similar syntactic, morphological, combinatorial, and prosodic properties.

Lexicographically, the existence of semantic classes means that lexical items comprising them

should receive similar lexicographic treatment, i.e., be described as belonging to the same *lexicographic type* (Apresjan 1995, 2000). That means not only streamlining their definitions in the dictionary, using *pro forma* (Atkins, Rundell 2008), or unified definition templates for words belonging to the same semantic class (e.g. animals, fruits, devices), but also describing their other linguistic properties uniformly. However, in order to fulfil that natural requirement, a lexicographer should possess pre-existing knowledge about all the semantic classes in language, along with their composition, structure, and linguistic peculiarities. For example, the majority of speech verbs as a semantic class in Russian share the following properties that should be reflected in a dictionary consistently:

- *semantic properties*: definition where the component 'to say' or 'to speak' is the topmost predicate in the assertive part of the semantic structure; actantial semantic structure with semantic roles of the speaker, addressee, topic and content of the utterance;
- *syntactic properties*: actantial syntactic structure with speaker expressed as nominative, content expressed as accusative or 'that'-proposition, addressee expressed as dative and topic expressed as prepositional;
- *morphological properties*: derived verbal nouns with resultative meaning ('talk', 'request', 'murmur'), but not with agentive meaning (*'talker', *'requester', *murmurer).

However, while certain classes are known relatively better (e.g. verbs of speech), due to their linguistic prominence, many other semantic classes are not even identified as such at the current stage of semantic and lexicographic development. That is precisely the area where active dictionary turned into a searchable database could prove to be a useful tool of research, since it provides structured and formalized multi-level linguistic information on every lexical item.

1.2 Sample Entry in ADR

A brief illustration on the types of information and structure of an ADR entry may be supplied by the Russian polysemous verb *bit'* 'to beat' in two of its most contrasted meanings, while the total number of its meanings in ADR is twenty six (by Valentina Apresjan in Apresjan et al. 2014).

As many highly polysemous verbs in Russian, *bit'* 'to beat' displays the effect of gradual semantic bleaching, accompanied by syntactic, morphological and prosodic change in its figurative senses as opposed to its direct senses (cf. the notion of colligation, or grammatical change signalizing change in meaning (Atkins, Rundell 2008)). This contrast is most pronounced in "physical" senses of *bit'* vs. its "light verb" senses, such as *bit' 2* 'to beat up as punishment' vs. *bit' 7.3* 'to chime (about clocks)'. The most striking differences are observed in the semantic and syntactic actantial structure, morphological properties, prosodic properties and derived nouns. Below are abridged excerpts from the dictionary entries illustrating the major differences between the senses of *bit'*. For the sake of convenience, English glosses are used instead of Russian examples and definitions.

BIT' 2 **'**to beat smb.', imperfective; perfective *izbit'* 'to beat smb. up'.
Examples: 'to beat smb. in the face'; 'to beat smb.'s head against the wall'; 'to beat smb. with a belt'; 'to beat smb. for theft'.
Meaning: **'**A1 is beating A2 in A3 with A4 for A5' = 'A person A1 is repeatedly striking the body part A3 of the person or animal A2 in abrupt and strong movements, using an instrument or body part A4 in order to cause A2 pain or physical damage, sometimes in punishment for transgression A5'.
Morphology: in imperfective usually denotes a process: 'He beats-IMPERF-PRESENT a boy' means 'He is beating a boy now'.

Government pattern (syntactic actants expressing semantic actants):
Model 1:
A1 Nominative: 'The father-NOM is beating'.
A2 Accusative: 'to beat the thief-ACC'.
A3 *po* 'on' Dative: 'to beat on the head-DAT'.
v 'in' Accusative: 'to beat in the stomach-ACC'.
pod 'under' Accusative: 'to beat under the shoulder blade-ACC'.
Instrumental: 'to beat one's head-INSTR against the wall'.
A4 Instrumental: 'to beat with one's feet-INSTR'; 'to beat with a whip-INSTR'.
o 'against' Accusative: 'to beat one's head against the wall-ACC'.
A5 *za* 'for' Accusative: 'to beat for disobedience'.
Model 2:
A1 Nominative: 'The father-NOM is beating'.
A3 Accusative: 'to beat one's face'.
A2 Dative: 'to beat to Ivan-DAT his face'.
A5 *za* 'for' Accusative: 'to beat for cheek'.
Collocations: 'to beat cruelly <painfully>'; 'to beat to death'; 'to beat with one's fists'.
Derivatives: bit'je 'beating'; *poboi* 'beatings, blows'.

BIT' 7.3 'to strike', imperfective; perfective *probit'* 'to strike a certain hour'.
Examples: 'The clock is striking midnight'; 'The Kremlin clock chimes twelve times'.
Meaning: 'A1 is striking A2' = 'A clock or pendulum A1 at regular intervals produces a number A2 of musical noises that signify the onset of time A2'.
Prosody: when A2 is expressed, the sentence usually does not bear phrasal stress: 'The clock is striking MIDNIGHT', but not *'The clock is STRIKING midnight'.
Morphology: in imperfective usually denotes a result: 'The clock strikes-IMPERF-PRESENT midnight' means 'The clock has struck midnight'.
Government pattern (syntactic actants expressing semantic actants):
A1 Nominative: 'The clock-NOM is striking'.
A2 Accusative: 'to strike midnight-ACC'; 'to strike twelve-ACC'.
Derivatives: boj 'chime, striking of a clock'.

As can be easily seen, these two senses of Russian *bit*' differ in their semantic actants and semantic roles (five actants for *bit' 2* vs. two actants for *bit' 7.3*); in their syntactic expression (very diverse syntactic expression with possible change of diathesis for *bit' 2* vs. very scarce syntactic expression for *bit' 7.3*); in their morphology (different forms of perfective; different interpretations of imperfective); in their prosody (reduced ability to bear phrasal stress in *bit' 7.3*). Thus, semantic bleaching is paralleled by reduction or change in all other properties of a lexical item, and ADR allows one to observe these correlations, as it provides multilevel linguistic information, and not merely definitions and examples.

2 ADR as an Online Electronic Database

2.1 Database Structure

All the data for the database is taken from the original ready-to-print .RTF-files by parsing them with regular expressions. This process is not complicated as the dictionary is a well-structured system

itself. As a result, the structure of the ADR non-relational database almost identically repeats the structure of its printed version where different parts of one dictionary entry are represented as key-value pairs. Each entry therefore is stored in separate JSON-format file. The files are indexed by the Java-based full-text search engine which acts as a separate server, but is connected to the main application server written in Python programming language. This technology stack appears to be the most suitable for functioning with advanced search queries and online application rendering with Flask framework. Each headword (word with all of its meanings, as opposed to lexeme – word in one of its meanings) has its unique ID, which is used to generate links all over the dictionary. For example, word X is present as a synonym or a collocate in the Dictionary Entry of the word Y, user could get to Dictionary Entry of the word X by simply clicking on the word X.

Further development plan includes building an editor connected right to the database for ADR authors, therefore preliminary text-files parsing will not be needed. In addition, the data structure is made flexible and expanding its functionality always remains possible.

2.2 Advanced Search Capabilities

As stated above, advanced search features are practical in performing language research. Besides the usual one-word search, we have also introduced three additional types of search.

2.2.1 Zone Search

This is a simple text search in one selected zone of the 13 dictionary zones (Example, Definition, Comment, Government Pattern, Construction, Collocations, Literary Illustration, Synonyms, Antonyms, Analogues, Conversives, Derivatives and Phraseology) in all of the Dictionary Entries of ADR. Although it is a simple technical decision, it may be practical in many ways. Consider the example in Section 3, where the search for "used in negative sentences" in the Comment zone allows one to extract in one click all the negative polarity items described in the dictionary. To a language learner, besides the obvious convenience of using an electronic tool instead of a paper dictionary, this type of search may provide unexpected benefits. For example, searching by a semantic component (by a word in the zone of the Definition), the user can extract words (s)he does not know or remember: e.g. the search for 'speak' in the definition would return all verbs of speech, including more peripheral ones, such as '*besedovat'* 'to converse without haste and with pleasure', '*boltat'* 'to chat', etc.

2.2.2 Search by Government Pattern

In this section the user can select Part of Speech of the word, Actant number (A1, A2 etc.) or the total quantity of Actants, Actant syntactic expression (NP in the form of a particular case, PP of a certain type, infinitive, gerund, a particular type of sentence) and Actant Semantic Role.[1] This type of search allows one to study the correlations between semantic and syntactic actants, including the less obvious ones, such as blending (two semantic actants are expressed as one syntactic entity); change of diathesis (a predicate has more than one government pattern); inexpressible semantic actant (no possible syntactic expression for a semantic actant). It also provides the opportunity to analyze language-specific syntactic expression of different semantic roles. Finally, this search option helps establish semantic classes and lexicographic types with their unique sets of semantic and syntactic

[1] Semantic roles are not included in the original version of the dictionary and are currently in the process of assignment, using the inventory of semantic roles (Apresjan et al. 2005).

features, as well as advance general linguistic knowledge on semantics-syntax interface. To a language learner, it provides helpful information on word usage – which noun case or verb form to choose, which preposition to use.

2.2.3 Search by Semantic Tag

Each ADR lexeme is assigned different kinds of semantic tags (such as 'action', 'good', 'building', etc.), thus, the user can choose one or many semantic tags and receive a wordlist of all relevant lexemes in ADR. [2] Again, this option allows one to juxtapose items belonging to a particular semantic class and lexicographic type as a way of studying their common properties or correct inconsistencies in their description. To a language learner, it can be a useful instrument of self-instruction where the user can study words of a certain thematic class – buildings, fruits, animals, emotions, etc.

2.3 Interface

The interface is also one of the major aspects of a good modern research tool. We paid special attention to the simplicity when designing ADR Online to make its use as easy as possible. Consequently, lightweight and clean appearance of the resource helps get the researcher the maximum of its functionality. The menu and main search form are always fixed on the left side, while the bigger right part accommodates all the other content including dictionary entries and advanced search properties (see figure 1 and 2). Our interface is made mobile-friendly and therefore it can be easily used from any device.

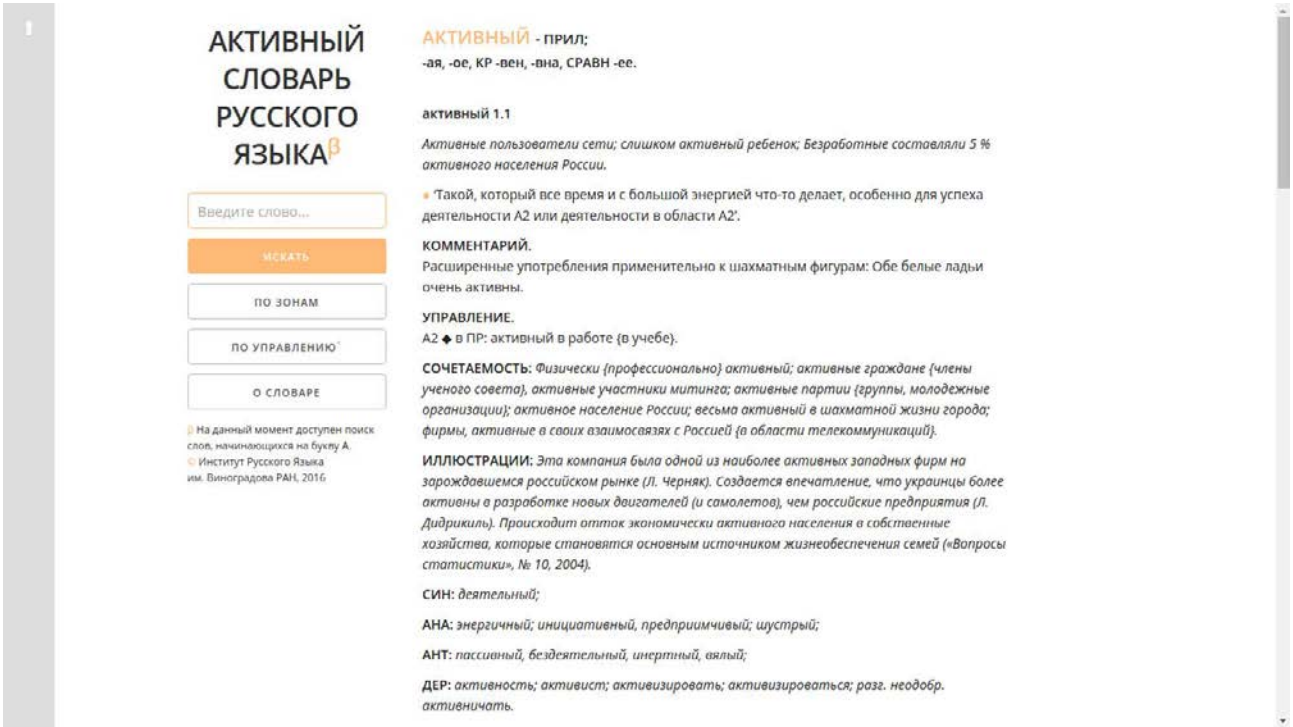

Figure 1: Dictionary entry page.

[2] Semantic tags are not included in the original version of the dictionary and are currently in the process of assignment, using the inventory of semantic tags from (Apresjan et al. 2005).

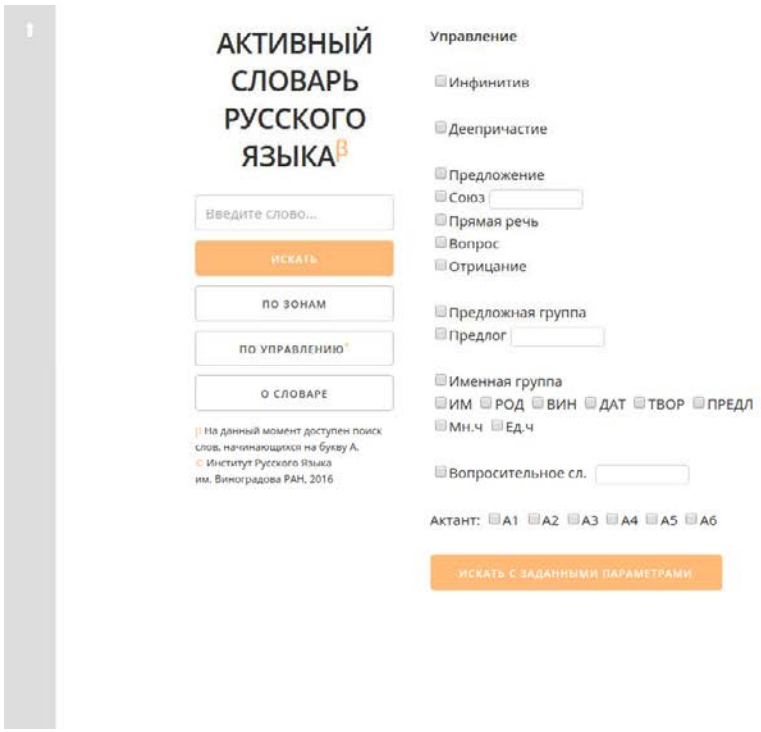

Figure 2: Search by government pattern page.

3 Sample Use of ADR as an Instrument of Linguistic Research Figures

As stated above, even simple zone search of such a dictionary as ADR Online, with its wealth of structured linguistic information, can be fruitful and produce theoretically unexpected results. The semantic class in question are the so-called negative polarity items, or words which are restricted in their usage to non-veridical contexts, such as negative, interrogative, and conditional sentences, downward entailing quantifiers (such as *few*), modals and some others (Giannakidou 1998).

A classical negative polarity item is the pronoun *any* and its derivatives, such as *anything, anyone,* etc. While it cannot occur in affirmative sentences (cf. the ungrammatical **He brought any books*), negative, interrogative, downward entailing and other non-indicative (non-veridical) contents license its use: *He didn't bring any books*; *Did he bring any books? Few professors brought any books.* Although negative polarity items are quite a popular topic in theoretical semantics, usually their consideration is limited to *any*, as well as certain phraseological expressions, such as *(not) to stir one's finger, (not) to do a stroke of work, (not) give a second thought, (not) give a rat's tail about* and similar idioms.

However, ADR Online demonstrates that the phenomenon of negative polarity is much more widespread than previously thought, and that many lexical items are capable of developing negatively polarized meanings (cf. notes to this effect in Apresjan 2012). Although a profound analysis of this semantic class remains a topic for future study, even a brief search reveals considerable material that expands our understanding of negative polarity.

ADR, as an active dictionary type, provides not only information on the meanings of the words but also on the typical conditions of their use – syntactic, pragmatic, lexical. For items that prefer negative, interrogative or conditional contexts, this information is provided in the Comment zone, after the definition. Therefore, search for "used in negative sentences" in the Comment zone returns the desired lexical items. Their analysis yields the following major semantic classes of words that tend to develop negative polarity usages or meanings:

- verbs with the semantic component of motion, such as *vorochat'sja* 'to roll slowly', *v'jexat'* 'to drive in', *vtykat'* 'to stick in', *vylezat'* 'to climb out, to extricate oneself';
- verbs with the semantic component of physical impact *brat'* 'to take', *vyderzhat'* 'to bear', *vynosit'* 'to stand', 'to tolerate'

Thus, the above-mentioned verbs develop the following negative polarity meanings:

- 'to roll slowly' → 'to be able to move'

(1a) *U nego jazyk ne vorochalsja* 'He was unable to move his tongue', cf. the awkward phrase
(1b) ??*U nego jazyk vorochalsja* 'He was able to move his tongue'

- 'to drive in', 'to stick' → 'to be able to understand'

(2a) *On nikak ne mog vjexat' v to, chto ja govoril* 'He was unable to get what I was saying'
(2b) ?*On vjexal v to, chto ja govoril* 'He got what I was saying'

(2c) *On ne vtykaet* 'He is unable to understand (anything)'
(2d) ??*On vtykaet* 'He is able to understand (everything)'

- 'to climb out, to extricate oneself' → 'to get out, to leave'

(3a) *On nedeljami ne vylezaet iz svoej kvartiry* 'He doesn't leave his apartment for weeks'
(3b) ??*On kazhdyj den' vylezaet iz svoej kvartiry* 'He leaves his apartment every day'

- 'to take' → 'to be able to get through'

(4a) *Lopata ne brala merzluju zemlju* 'The shovel couldn't cut through the frozen soil'
(4b) ?*Lopata brala merzluju zemlju* 'The shovel could cut through the frozen soil'

- 'to bare', 'to stand' → 'to tolerate', 'to outbrave'

(5a) *On ne vynosit podlosti* 'He doesn't tolerate meanness'
(5b) ??*On vynosit podlost'* 'He tolerates meanness'

(5c) *Ona ne vyderzhala i rasplakalas'* 'She broke down and dissolved in tears'
(5d) ??*Ona vyderzhala i ne rasplakalas'* 'She didn't break down and didn't dissolve in tears'

It is not difficult to notice what unites all these verbs: either in their direct or figurative meanings they all contain an indication of *difficulty* and *obstruction* that needs to be overcome, cf. the senses that reflect this idea: 'slowly', 'to extricate', 'to be able', 'to outbrave'. This indication of effort seems to be a factor that plays a role in negative polarization. Indeed, it is quite logical that actions requiring greater effort are less controllable and therefore less likely to succeed than actions that are easy to perform. These semantic components are preserved and even strengthened in metaphorical meanings, thus, it is only natural that verbs that develop such meanings are more likely to appear in negative and other non-veridical contexts than in indicative ones.

The observed semantic tendency allows one not only to explain but also to predict what words might develop negatively polarized senses. Consider e.g. the motion following verbs: *povorachivat'sja* 'to turn', *podnimat'sja* 'to lift', *vysovyvat'sja* 'to stick out'. In their figurative meanings they are negatively polarized; as in *U nego jazyk ne povorachivaetsja vozrazit* 'He does not have the heart to object', but not *'He has the heart to object'; *U nego ryka ne podnimaetsja jejo udarit'* 'He cannot bring himself to hit her', but not *'He brought himself to hit her'; *On nikogda ne vysovyvaetsja na nashix sobraniax* 'He never sticks his head out at our meetings; He always keeps a low profile', but not *'He always sticks his head out at our meetings, He always keeps a high profile'.

The above is but one example of how electronic lexicography can advance linguistic theory; with ADR Online more are expected to follow.

4 References

Apresjan J.D. (1995). Integral description of language and systematic lexicography [Integral'noe opisanie jazyka i sistemnaja leksikorafija]. Moscow. (In Russian).

Apresjan J.D. (2000). Systematic lexicography. Translated by Kevin Windle. Oxford University Press.

Apresjan J.D. (2012). Verbal grammar in the Russian Active Dictionary [Grammatika glagola v Aktivnom slovare russkogo jazyka]. In Meanings, texts and other fascinating plots [Smysly, teksty I drugie zaxvatyvajushchie sjuzhety]. A collection of articles in honor of Igor Mel'čuk. Moscow.

Apresjan J. D., Boguslavskij I. M., Iomdin B. L., Iomdin L. L., Sannikov A. V., Sannikov V. G. and Sizov L. L. (2005). Syntactically and semantically annotated corpus of Russian language: Present state and perspectives [Sintaksicheski i semanticheski annotirovannyj korpus russkogo jazyka: sovremennoe sostojanie i perspektivy], National Corpus of Russian Language: 2003–2005 [Natsional'nyj korpus russkogo jazyka: 2003–2005], pp. 193–214. (In Russian).

Apresjan J.D., Apresjan V.J., Babaeva E.E., Boguslavskaja O.J., Galaktionova I.V., Glovinskaja M.J., Iomdin B.L., Krylova T.V., Levontina I.B., Ptenstova A.V., Sannikov A.V., Uryson E.V. Ed: Apresjan J.D. (2014). Active Dictionary of the Russian Language [Aktivnyj slovar' russkogo jazyka]. Moscow.

Atkins B.T.S., Rundell M. (2008). The Oxford Guide to Practical Lexicography. Oxford University Press.

Giannakidou, A. (1998). Polarity Sensitivity as (Non)veridical Dependency. John Benjamins, Amsterdam-Philadelphia.

Acknowledgements

This project was partly funded by the following grant: Russian Humanities Fund № 16-04-00302 "The third issue of the Active Dictionary of Russian"; 2016-2018.