Development of the Lexicon of Argumentation Indicators

Irina Kononenko(✉) and Elena Sidorova

A.P. Ershov Institute of Informatics Systems SB RAS,
Novosibirsk 630090, Russia
irina_k@cn.ru, lsidorova@iis.nsk.su

Abstract. The paper presents the results of a preliminary analysis of the argumentation indicators observed in the corpus of popular science texts in Russian. Main pragmatic aspects of the argumentation signaled by discursive indicators are outlined. The classification of indicators takes into account pragmatic meaning and the type of language means used. Special attention is paid to insufficiently studied indicator constructions and classes of their core content words. We consider constructions with verbs and nouns of mental state, speech, inference, and mental impact. The process of creating a lexicon of argumentation indicators is described. Indicators are presented in the form of lexical units and lexical-grammatical patterns, which are automatically generated from annotated text fragments and can be manually corrected by the expert. The pattern description language allows to represent grammatical and semantic constraints, nested constructs, alternatives, and discontinuity. The lexicon of indicators will be used for automatic annotation of argument indicators in unannotated text, as well as for experiments in argument mining.

Keywords: Argumentation indicator · Lexical-grammatical pattern · Argumentative annotation · Popular science discourse · Text corpus

1 Introduction

In the last decade, an interdisciplinary field of research has been actively developing at the intersection of philosophy, psycholinguistics and computational linguistics. Its purpose is to create models of argumentation for various types and genres of discourse and automatically identify and extract argument components and structure including premises and conclusions, and the relations between them based on typical argumentation schemes. The main prerequisite for the development of this area is the creation of annotated corpora, in which textual fragments are matched with components of argumentative structures and relations between them.

So far, there exist only a few resources with annotated argumentation structures over monologue texts, mainly for the English language. The best known is AIFdb[1], the

[1] AraucariaDB, http://corpora.aifdb.org/araucaria, last accessed 2019/04/30.

© Springer Nature Switzerland AG 2019
S. O. Kuznetsov and A. I. Panov (Eds.): RCAI 2019, CCIS 1093, pp. 154–168, 2019.
https://doi.org/10.1007/978-3-030-30763-9_13

former Araucaria corpus [1], which contains news articles, records of parliamentary and political online debates. Resources are created in German: University of Darmstadt Corpus[2] includes subcorpora of student essays [2], news texts and scientific articles; the Potsdam corpus[3] contains a small set of microtexts on a given topic, later translated into English [3]. There exist projects for some other languages (Italian, Greek, Chinese). As for the Russian language, such resources, as far as we know, do not yet exist. In most cases, corpus annotation includes text segmentation with highlighting of argumentation units, markup of roles (premise, conclusion) and relations (support, attack), without matching the argumentation schemes on which the reasoning is based. An exception is Araucaria, where argumentative structure annotation is related to particular argumentation scheme based on the theory of Walton [4].

The proposed work was performed as part of an on-going research project aimed at creation of an argumentation annotated corpus for the Russian language. A popular science discourse that is not presented in well-known argumentatively annotated corpora is being studied. Popular science discourse is defined as a way of transmitting scientific knowledge or innovation projects by the author-scientist (or a journalist as an intermediary) for their understanding by a mass audience. The corpus of popular science online articles on linguistic topics has been selected with the help of catalogs of Russian search engines Yandex and Rambler. Corpus includes about 70 texts with an average volume of 1057 words (minimum – 167 words, maximum – 4094 words), with no restrictions on the subject, structure, and the type of presentation. Some articles are transcripts of oral presentation, interviews, etc.

The texts are annotated manually based on the argumentation model developed by the project participants. An important linguistic aspect of the process of arguments annotation is registration of argumentative indicators, which constitute keystones in the discourse, facilitating the identification and reconstruction of argumentative moves that are made in argumentative discussions and texts (see [5]). Argumentative indicators are language means (words, constructions) that serve as discourse clues in identifying the structure of argumentation: they help determine the presence of arguments in a given segment of text, reconstruct the connections between statements, relate the argument to a specific reasoning pattern (inference form expressing the relations of premises and conclusions).

The purpose of this study is to create a lexicon of argumentative indicators used in popular science discourse. The work outlines the preliminary results of the analysis of argumentative indicators selected in the corpus of popular science articles. The questions of their classification, structural features and methods of formal representation are discussed.

[2] TU Darmstadt Homepage of Argumentation Mining, www.informatik.tu-darmstadt.de/ukp/research_6/research_areas/argumentation_mining, last accessed 2019/04/30.

[3] Potsdam corpus, http://angcl.ling.uni-potsdam.de/resources/argmicro.html, last accessed 2019/04/30.

2 Related Works

Discourse markers (discourse connectives) are usually considered as key indicators of discourse structure. They have been studied from various research perspectives. One of them is represented in Penn Discourse Treebank where discourse connectives are viewed as binary predicates that convey certain semantic relations and take propositions, events and states as their arguments PDTB [6]. PDTB annotation covers traditional functional words and phrases such as subordinating conjunctions (e.g. *when, because, as soon as*), coordinating conjunctions (*and, but, or*), adverbs (e.g. *instead, therefore*), prepositional phrases (e.g. *on the other hand*), etc.

T. van Dijk proposed classifying discourse connectives according to the type of relation they label: pragmatic connectives express the relation between speech acts, semantic connectives manifest the relations between the facts indicated in the text [7]. This difference corresponds to the opposition of subject matter and presentational relations in the Rhetorical structure theory [8]. Presentational rhetorical relations whose intended effect is to increase some inclination in the reader, such as the desire to act or the degree of positive regard for, belief in, or acceptance of the nucleus, overlap with argumentative discourse relations. The mapping of rhetorical discourse relations onto argumentative relations carried out in [9] confirms this pragmatic similarity. No wonder that first experiments in argumentation mining use the traditional functional lexicons as lexical indicators.

Stab and Gurevych [10] experimented with different types of features, including discourse markers from the PDTB annotation guidelines, to classify text units into the classes non-argumentative, major claim, claim, and premise. The PDTB markers appeared to be not helpful for discriminating between argumentative and non-argumentative text units, but they were useful to distinguish between the classes premise and claim. Eckle-Kohler et al. [11] present a study on the role of discourse markers in argumentative discourse on the material of German corpus, with arguments annotated according to the common claim-premise model of argumentation. They performed various statistical analyses regarding the discriminative nature of discourse markers for claims and premises. The experiments show that particular semantic groups of discourse markers are indicative of either claims or premises and constitute highly predictive features for discriminating between them.

The investigation of discourse relation signals given in [12] is more extensive, as it takes into account not only traditional discourse markers (e.g., *although, because, since, thus*), but also signals such as tense, lexical chains or punctuation, and their combinations. The authors of the project to create a corpus of rhetorical structures on the material of the Russian language[4] also consider a wide class of language expressions, including lexical items irrespective of their part of speech that can signal the presence of a rhetorical relation. Toldova et al. [13] consider not only functional words to be rhetoric relation markers. The markers include punctuation marks, prepositions, pronouns, speech verbs, etc. In the development of this approach on the example of causal relation indicators in [14] it is shown that, in addition to traditional functional

[4] Russian RST Discourse Treebank, https://linghub.ru/ru-rstreebank, last accessed 2019/04/30.

words, relation indicators are constructions based on the content words and provide informal specifications of some patterns that can be used for mining indicators in non-annotated text.

With regard to the indicators of the argumentation, the possibility of considering a wide class of language expressions that signal the use of specific reasoning schemes is demonstrated in the theoretical study [5], which also goes far beyond the functional classes of words. Considering the indicators of argumentation by analogy, the authors cite as an example constructions with significant words meaning analogy, comparison, similarity, and parallelism: *X can be compared to Z; X is similar to Z; X is the equivalent of Z; there are parallels (to be drawn) between X and Z; X reminds someone of Z.*

3 Information Model of Argument Annotation

An argument is a set of related statements used to prove a final statement (thesis, or conclusion). The structure of the argument highlights the statement-premise and the statement-conclusion connected by typed relations.

The structure of the argument can be represented as follows:

Argument = (Premise, Premise, …, Conclusion, Weight)
Conclusion = (Statement | Argument, Support | Attack, Weight)
Premise = (Statement, Role, Weight)
Statement = (Utterance, Source *, impl. | expl.)

The type of argumentation relation expresses whether a given argument is evidence (*Support*) or refutation (*Attack*) of a thesis-conclusion. The conclusion can be either an explicitly expressed statement or some other argument. Related statements may serve as premises, where each premise plays a specific *Role* in a typical reasoning scheme.

A statement represents a natural language formulated proposition (*Utterance*), which the annotator (expert) associates with the *Source* that is a text fragment. Usually the statement coincides with the source, except for the existing anaphoric references and ellipsis recovered by the annotator from the context. Thus, a statement is an interpretation of a text fragment. However, it is possible that the necessary statement-premise is not explicitly specified in the text. In the case of implied premise, its statement can be formulated by the expert on the basis of extratextual knowledge.

All elements in the structure of the argument are supplied with *Weight* – a measure of the persuasiveness of the proof given, which allows us to ultimately assess the strength of the author's argument as a whole.

The given argument representation model corresponds to the AIF model [15], which is currently accepted as a standard in analyzing argumentative structures and, in particular, is used in the Carneades system [16]. Since in this study we focused on investigation of different types of indicators used in the texts for entering arguments and their structural components, the argument model was supplied with additional parameters for annotating the argumentation indicators in the text.

Indicator = (Source, Type, Definition, Frequency)

On discovering an indicator, the expert marks up a corresponding text fragment (*Source*) and points out which pragmatic aspect (*Type*) of the argument is signaled by the indicator. Based on the analysis of the selected fragment, the structural (grammatical) type of the indicator is determined and its lexical-syntactic *Definition* formed, which allows automatic search for the indicator in the texts. The *Frequency* parameter determines how discriminative this indicator is for the selected aspect of the argument. Frequency in the annotated text corpus is calculated automatically.

Additionally, the markup system implements the requirement of maximum "similarity" between the statement and the source. To this end, the following recommendations were developed for experts who carry out annotation of argumentation.

When annotating an *Argument*, text fragments corresponding to the explicitly presented statements are marked up first. Each fragment can be a chain of sentences, a single sentence, clause or nominalization. Every fragment is regarded as if all its anaphoric references (including ellipses) were resolved. In case of anaphoric nominalization of a whole statement within the Argument, an antecedent statement is marked up. Then, a suitable type of reasoning scheme (argumentation scheme) is chosen, the selected statements are linked into a single Argument, and the necessary parameters of the premises and a conclusion are indicated in accordance with the specified scheme. If necessary, implicit statements are introduced.

Let's give an example of the Argument marked up in the text[5]:

(in Russian) *По-французски любовь – amour, что тоже имеет тайный смысл.* [*Звукосочетание "mr" в индоевропейском праязыке соответствовало всему, что связано со смертью.*] [*Звук 'a' до сих пор во многих языках употребляется как противопоставление.*] ***Поэтому*** [*«amour» – противопоставление смерти, то есть жизнь!*]//text 21

In French love - amour, which also has a secret meaning. [*The sound combination "mr" in the Indo-European proto-language corresponded to everything connected with death.*] [*The sound 'a' is still used as an opposition in many languages.*] ***Therefore*** [*«amour» is the opposition of death, that is, life!*]

In this example, the Argument consists of two premises and a conclusion. The word *поэтому 'therefore'* is an indicator of the conclusion of the Argument and of entire inference relation.

Note that the Argument does not always correspond to a continuous text fragment: between the conclusion and the premise there may be discourse units that are not related to this Argument (for example, Premise that supports the same Conclusion independently within another Argument), or irrelevant for argumentation (for example, explanations).

[5] Statements in the structure of the argument are presented in square brackets. The statement that presents the conclusion of the argument is underlined. Indicators are bold italic. After the fragment the source text is given.

4 Classification of Argumentation Indicators

Indicators of argumentation can be classified from different points of view: the pragmatic aspects of argumentation, the degree of grammaticalization, the semantics of the indicator's core word, the type of construction.

1. Pragmatic aspects of argumentation signaled by the indicator.
   - opinion and strength of the argument (degree of confidence);
   - inference relation between two statements;
   - role of the statement in the inference relation (Premise vs. Conclusion);
   - type of argumentative relation (Support vs. Attack);
   - structure of the argumentation (Multiple vs. Serial argumentation);
   - semantic-ontological relation which the typical reasoning scheme used in this case is based on.

In the following examples (1) and (2), the indicators *по-видимому 'seemingly'* and *специалисты предполагают, что 'experts suggest that'* present statements of the premise (2) and conclusion (1) as opinions with a certain weight. Indicators *поскольку 'since'* and *поэтому 'therefore'* with causal semantics explicitly indicate the presence of a relation of inference. In this case, the position of the marker in the segment indicates the role of the corresponding statement: *поскольку* introduces the Premise in (1), and *поэтому* introduces the Conclusion in (2). In both cases, the type of relation is Support. In (3) and (4), the indicators are based on predicates with the semantics of mental impact, *опровергать 'refute'* and *подтверждение 'confirmation'*, here the distribution of roles in the inference move is identified by the actant position.

(1) ***Поскольку*** [*в языках сибирских народов все еще сохранилась четкая связь с индейскими наречиями*], ***специалисты предполагают, что*** [*многие мигранты возвращались из Америки назад, в Сибирь*.]//text 02

Since [*in the languages of the Siberian peoples there is still a clear connection with Indian dialects*], ***experts suggest*** *that* [*many migrants returned from America back to Siberia*.]

In the example (1), the opinion of specialists expressed in the conclusion and marked by an indicator of opinion, which corresponds to a not very high weight (the degree of confidence of the mental predicate is relatively low), is supported by the premise marked by the indicator of the basis of the conclusion.

(2) [*Осознание своей идентичности, в том числе и языковой, по-видимому, является важным компонентом душевного равновесия.*] *Именно* ***поэтому*** [*всегда находятся те, кто наперекор современным тенденциям, а то и инстинкту самосохранения поддерживает и сохраняет языки*.] ***Тем более что*** [*знание родного языка совершенно не означает отказа от других, более востребованных.*]//text 68.

[*Awareness of one's identity, including linguistic identity, is* ***probably*** *an important component of mental equilibrium.*] *Just* ***for that reason*** [*there are always those who, contrary to modern trends and even to the instinct of self-preservation, maintain and preserve languages*.] ***All the more so that*** [*knowledge of the mother language does not mean refusal to speak other, more popular ones.*]

In the example (2), two arguments are shown that prove the same thesis independently of each other, while the indicator *тем более что 'all the more so that'* marks the second premise in the structure of Multiple argumentation.

(3) *Например,* ***поговаривают, что*** [*русских научили материться татары и монголы, а до ига, якобы, не знали на Руси ни одного ругательства*.] ***Однако*** *есть несколько* ***фактов, опровергающих это***. *Во-первых,* [*у кочевников не было обычая сквернословить.*]//text 29.

For example, ***they say that*** [*the Tatars and the Mongols taught Russians how to swear and before the yoke, allegedly, they did not know a single curse in Russia*.] ***However***, *there are several* ***facts that refute this***. *First,* [*nomads didn't have the habit of foul language.*]

(4) *Во-первых,* [*у кочевников не было обычая сквернословить*.] ***В подтверждение этому*** *— [записи итальянского путешественника Плано Карпини, посетившего центральную азию. Он отмечал, что у них бранные слова вообще отсутствуют в словаре.]//*text 29.

First, [*the nomads did not have the habit of foul language*.] ***In confirmation of this*** *—* [*the records of the Italian traveler Plano Carpini, who visited Central Asia. He noted that swear words were absent in their lexicon.*]

Examples (3) and (4) demonstrate Serial argumentation. In (3) an opinion is refuted by the following premise (Attack relation), and in (4) this premise is supported by the reasoning corresponding to the typical scheme "From the Knower": the subject makes a statement relating to the domain he is familiar with - therefore, this statement is true.

2. Primary and secondary indicators.
Toldova et al. in [14] proposed to divide the indicators of a causal rhetorical relation into two classes (primary vs. secondary) according to the degree of their grammaticalization: the primary connectors are functional words (including multi-word units) fixed in grammars and dictionaries, and the secondary ones are less studied constructions based on content lexemes of causal semantics. Examples from the corpus of popular science texts make it possible to draw similar conclusions regarding argumentation indicators. We consider two classes of language means used as indicators of argumentation:

– discursive connectors are well-known functional units, including multi-word units (prepositions, conjunctions, introductory words): *поэтому 'that is why', поскольку 'since', следовательно 'consequently', так как 'as', значит 'hence', тем более что 'all the more so that', например 'for example', в частности 'in particular'*, etc.;
– content words and indicator constructions including these words as their core components (see examples below).

3. Classification of indicators according to the semantics of the core content word.
Up to now the list of annotated content words which can serve as indicators or core words of indicator constructions is heterogeneous and far from complete. These words are mainly verbs and nouns of the following lexical-semantic classes:

- **mental state** *считать 'to believe', предполагать 'to suppose', убежден 'be convinced', мнение 'opinion', точка зрения 'viewpoint'*;

- **mental impact** *доказывать 'to prove', опровергать 'to refute',подтверждать 'to confirm', свидетельствовать 'to indicate'*;
- **inference** *следовать 'to follow/result', получается 'it follows that', выходит 'it follows that', выходить 'to follow/result', получаться 'to follow/result', выводить 'to conclude/infer', следствие 'consequence', вывод 'conclusion'*;
- **conflict** *противоречить 'to contradict', противоречие 'controversy'*;
- **intellectual *activity*** *обнаружить 'to discover', выяснить 'to find out', выявить 'to reveal'*;
- **speech activity** *говорить 'to talk', сообщать 'to report', утверждать 'to state'*;
- **justification** *аргумент 'argument', доказательство 'proof', обоснование 'basis', свидетельство 'evidence', подтверждение 'confirmation'*;
- **information** *факт 'fact', пример 'example'*;
- **intellectual product** *тезис 'thesis', гипотеза 'hypothesis', теория 'theory'*;
- **speech product** *сообщение 'message', слово 'word'*;
- **expert** *ученый 'scientist', специалист 'specialist', лингвист 'linguist', философ 'philosopher'*.

4. Types of constructions for secondary indicators.

On the basis of speech and mental predicates, predicates of inference and mental impact, complex indicators of argumentation are formed. In addition to the core word, they can include markers of actant positions, for example, the conjunction *что 'that'* and the correlative pronoun construction *то, что 'the fact that'* for sentential actants, anaphoric and cataphoric elements such as the demonstrative pronoun *это/этот 'this'*, adverb *отсюда 'hence'*, the relative pronoun that 'what'. Examples of constructions under consideration are as follows:

– constructions with verbs of inference and mental impact
  из…следует, что 'from… it follows that'
  это…доказывает, что 'this… proves that'
  эти…свидетельствуют о том, что 'these… indicate that'
– verbal constructions of direct or indirect speech or opinion with the speech or mental verb and the «expert» class word in the subject position
  ученые… утверждают: "…" 'scientists…assert: "…" '
  литератор… заметил, что 'literary scholar…noted that'
– light verb constructions with nouns
  примером…является 'example …is'
  аргумент был такой 'argument…was as follows'
  приводит… аргумент в пользу этого, что 'give an argument in favour of this'
  отсюда … сделан… вывод о том, что 'come to a conclusion that'
– prepositional noun phrases
  в подтверждение этому 'in confirmation of this'
  на этом/таком основании 'on this/that basis'
  на следующем основании 'on the following ground'
  по мнению/словам 'according to smb'

5 Technological Aspects of Building a Lexicon of Argumentation Indicators

To support the development of lexicon of indicators, it is essential to provide the researcher with the necessary automation tools. In Fig. 1 the main stages of the process of creating and researching indicators are presented.

It is assumed that the process of argument annotation is accompanied by marking up argumentation indicators found out by the annotator. After a text fragment associated with the indicator is selected, a formal description of the indicator is automatically generated and added to the lexicon. This description is presented to the expert for validation and correction. Automated procedures carried out by the expert are supported by the appropriate software components.

Consider this process in more detail.

1. Selection of a text fragment corresponding to the indicator occurs together with annotation of the argument and its components. Analysis of the structure of the argument and the role of the indicator within this structure complement each other and facilitate annotator's work. The indicator annotation involves specification of the fragment boundaries (possibly with gaps) and selection of the argumentation aspect(s) signaled by the indicator.
2. Based on the selected fragment, it is necessary to specify a formal representation of the indicator in order to ensure automatic search of the indicator in the text, taking into account the variability of its presentation. At this stage, the text fragment is divided into elementary components (graphematic analysis), words are lemmatized, word combinations (phrases) are generated and normalized.
3. As the examples in paragraph 4 show, indicators are not only lexical units (single- or multi-word units), but also constructions, which can be formally represented by means of lexical-grammatical patterns. Automatically generated pattern allows for the lexical composition of the construction (lexical units in the normalized form), punctuation marks, gaps, and the boundaries of the indicator.
4. At the next stage, the obtained formal description is matched against the corpus and search results are displayed in the form of a concordance. Based on the study of the indicator's occurrences, the expert concludes whether the formal description is correct.
5. The expert can correct indicator description as appropriate: generalize individual lexical units to lexical-semantic classes, resolve ambiguities, specify grammatical features of words and phrases within structures (to ensure coordination or government), create lists of alternatives and indicate the boundaries of the construction.
6. The resulting lexical units and patterns approved by the expert are supplied with the necessary grammatical and argumentative features and introduced into the information retrieval lexicon, which provides search and automatic annotation of indicators in the texts of the corpus. This, on the one hand, removes the need to re-annotate indicators manually, and, on the other hand, signals the possible presence of argumentation in unannotated texts or the need to refine previously marked up arguments.

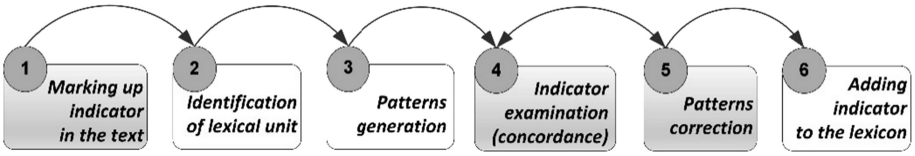

Fig. 1. The main stages of the development of lexicon of indicators (blocks with a light background correspond to fully automatic procedures, blocks with a dark background represent procedures carried out by an expert).

5.1 Indicator Pattern Generation

Indicators of argumentation can be classified from different points of view: the pragmatic aspects of argumentation, the degree of grammaticalization, the semantics of the indicator's core word, the type of construction.

The analysis of text fragments marked up as indicators is carried out using the Klan system [17]. Extraction of lexical units from a text fragment is not as obvious a task as it might seem. The paper [18] describes the emerging problems and gives a linguistic classification of errors. Most of the errors in the extraction of lexical units are related to the ambiguity and/or incorrect prediction of single words and the incorrectness and/ or incompleteness of the construction of word combinations.

The process of indicator pattern generation includes the following steps:

a. graphematic analysis, which provides for tokenization and selection of non-textual elements (numerical data, symbols, etc.),
b. lexical and morphological analysis (lemmatization, determination of lexical and grammatical features, paradigm representation, normalization),
c. identification of word combinations (based on predefined grammatical models and normalization),
d. generation of template (s) with a simple structure in the form of a chain of lexical units and punctuation marks:

   так, например 'thus, for example': [*так*, s/, , *например*]

e. for discontinuous fragments, introduction of structural constraints into the pattern description (distant context and pattern boundaries)

   если ..., то 'if...then': [begin: *если*, s/, , end: *то*]

f. analysis of pattern composition and ascription of grammatical features (for example, if the form of the indicator is fixed during annotation):

   в подтверждение 'in confirmation': [*в*, *подтверждение* <acc, nom, sing>]

g. analysis of the set of patterns and specification of formal description of pattern using compression procedures, such as introduction of alternatives, inclusion of references to other patterns, generalization and combination of patterns:

это 'it' or *этот 'this'*: [*это* | *этот*]

Thus, several types of structural organization of the formal description of indicators and their components can be distinguished.

Indicators with a simple structure include single- or multi-word functional and content units (inference predicates, speech and mental predicates, etc.).

Complex constructs described using patterns include simple chains (a chain of lexemes and punctuation marks), chains with grammatical constraints (prepositional phrases, verbal constructions, etc.) and discontinuous constructions.

Among the indicators with complex structural organization are the following:

– constructs combining distant context and grammatical constraints, for example, prepositional noun phrases

на ... основании 'on...ground'
[begin: *на*, end: *основание* <gen,sing>]

including auxiliary constructs with imposed grammatical constraints

согласно... теории/исследованиям/гипотезе
'according to...theory/research/ hypothesis'
research_group = [*исследование* | *теория* | *гипотеза*]
[begin: *согласно*, end: research_group<dat>]

– constructs with elements defined by their lexical-semantic classes:

это...доказывает, что 'this... proves that'
mental_impact_that = [w/<Sem:mental_impact>, s/, , *что*]
[begin: *это*, end: mental_impact_that]
из ...следует, что 'from... it follows that'
inference_ that = [w/<Sem:inference>, s/, , *что*]
[begin: *из*, end: inference_that]
'литератор... заметил, что' 'literary scholar...noted that'
speech_activity_that = [w/<Sem:speech_activity>, s/, , *что*]
[begin: <Sem:expert>, end: speech_activity_that]

– constructs with multiple gaps (distant contexts):

из... сделан... вывод о том, что 'from...come to ...conclusion that'
concl_that = [*вывод*, *о*, *том*, s/, , *что*]
[begin: *из*, w/<Sem:light_verb>, end: concl_that]

Correct and complete description of indicator in accordance with annotated fragment is not always obtained as a result of automatic template generation. The same goes for lexical-semantic class identification in case of generalization. Manual correction and adjustment of the formal representation of indicator is required based on the examination of its contexts and use in various types of arguments.

5.2 Analysis of Indicator Structure

The traditional tool for the study of linguistic phenomena is concordance, which displays a listing of immediate and extended contexts of lexical units in the text corpus. The advanced implementation of searching and concordancing, in addition to lexical units, provides contexts of pattern descriptions, with support for output filtering in accordance with specified criteria (for example, argumentation features). This functionality greatly increases the possibilities for research.

The goal of indicator examination carried out by an expert is to ensure the accuracy of the generated formal descriptions, as well as to expand the lexicon by identifying and merging indicators similar in structure and generalizing lexical units to lexical-semantic classes. The pattern description language has the necessary capabilities, such as means for representing grammatical and semantic constraints, nested constructs, alternatives, and discontinuity.

Let us consider the process of indicator patterns on the example of the "From the Expert" reasoning scheme commonly used in the popular science texts.

[[*«Достичь этого помогают гласные»*], - *добавляет канадский исследователь Сэм Мэглио (Sam Maglio)*], *один из авторов новой работы.//*text 01

[[*"Vowels help achieve this"*] *adds Canadian researcher Sam Maglio*], *one of the authors of the new work.*

In this example, the construction of direct speech is used, with the speech predicate and the «expert» class word in the subject position. This construction is generally recognized as a sign of argumentation. Thus, the annotator marked up the following text fragment as an indicator:

« ... » .. добавляет .. исследователь '« ... » .. adds .. researcher'

Based on this fragment, it is necessary to create a formal representation of the indicator. When generating a pattern, you can apply different strategies for forming its composition. For example, in this case the following pattern variants will be automatically generated:

– presentation of exact wordform with the help of grammatical features:

x = [begin: «, end: »]
y1 = [begin: x, *добавлять*<act,3pers,pres,sing>, end: *исследователь*<nom, sing>]

– presentation of all forms (normalization):

y2 = [begin: x, *добавлять*, end: *исследователь*]

– generation by grammatical model:

y3 = [begin: x, *добавлять*, end: *исследователь* <nom>]
y4 = [begin: x, *исследователь*<nom>, end: *добавлять*]

– specification of lexical-semantic class (with or without grammatical features):

y5 = [begin: x, w/<Sem: speech_activity>, end: w/<Sem: expert>], etc.

Determining the best strategy for each specific several types of indicator is one of the objectives of the study.

To expand and generalize the lexical composition of the generated pattern, the expert performs the following steps:

- considers the possibility of generalization of the core words by specifying their lexical-semantic classes,
- creates auxiliary patterns with alternatives,
- checks the generalization hypothesis using concordance,
- corrects and validates the indicator by checking all its occurrences in the corpus.

There are more than 300 occurrences of the «expert» class words: *исследователь 'researcher'* (40), *ученый 'scientist'* (119), *специалист 'specialist'* (17), *эксперт 'expert'* (6), *лингвист 'linguist'* (98), *филолог 'philologist'* (7), *антрополог 'anthropologist'* (2), *археолог 'archeologist'* (8), *профессор 'professor'* (15), *физик 'physicist'* (3), etc. The concordance listing shows that contexts of these words include the following lexical markers of argumentation: *добавлять 'add', пояснять 'explain', признавать 'admit', отмечать 'note', сообщать 'report', подытоживать 'summarize', резюмировать 'sum up'*, etc. These words were grouped into the lexical-semantic class «speech_activity» to be used in final patterns.

As a result of the correction carried out by the expert, there are patterns that describe a whole class of situations:

quote_l = [“|«] quote_r = [”|»] DS = [begin: quote_l, end: quote_r]
Expert = [w/<expert>] | [ph/<expert>]
DSC1 = [begin: DS, w/<speech><V, past|pres>, end: Expert<N, nom>]
DSC2 = [begin: Expert<N, nom>, w/<speech><V, past|pres>, end: DS]

Search in the corpus shows that the construction corresponding to this pattern appears 7 times and 6 of these occurrences indicate the presence of the “From the Expert” argumentation.

Another example of a complex pattern corresponds to the indicator used in the “From the Sign” argumentation scheme. The pattern represents a construction with verb of mental impact and anaphoric element in the actant position.

Это *открытие также* ***доказывает, что*** [*переселение народов из центральной Азии в северную Америку 13 000 лет назад, возможно, было не окончательным.*]//text 02

This *discovery also* ***proves that*** [*the migration of peoples from Central Asia to North America 13,000 years ago may not have been final.*]

to_chto = [s/ ,, *что*] | [*то*, s/, , *что*] | [*того*, s/, , *что*]
anaph_this = [*это* | *этот* | *такой*]
Proof = [begin: anaph_this, w/<caus_ment><V, pres>, end: to_chto]

The above examples demonstrate the technique of developing formal descriptions of indicators, including automatic generation and manual correction procedures.

6 Conclusion

The paper presents the results of a preliminary analysis of the argumentation indicators observed in the process of annotation of popular science texts in Russian. Corpus examples show main pragmatic aspects of the argumentation signaled by discursive indicators. Along with pragmatic meaning, the classification of indicators takes into account the type of language means used. Special attention is paid to insufficiently studied indicator constructions and classes of their core content words. We consider constructions with verbs and nouns of mental state, speech, inference, and mental impact.

The argumentation indicators are presented in the form of lexical units and lexical-grammatical patterns, which are automatically generated from annotated text fragments and can be manually corrected by the expert. The lexicon of indicators is planned to be used for automatic annotation of argument indicators in unannotated text, as well as for experiments in argument mining.

The process of argumentative annotation of the popular science corpus is ongoing. Upon completion of the work, the pilot version of the annotated corpus will be available in the open access. We assume that in the future the scope of the research will expand and cover new classes of content words and corresponding constructions. In particular, one can expect a significant expansion of the spectrum of indicators due to the semantic-ontological relations on which typical argumentation schemes are based.

Acknowledgments. The research has been supported by Russian Foundation for Basic Research (Grant No. 18-00-01376 (18-00-00889)).

References

1. Reed, C., Mochales Palau, R., Rowe, G., Moens, M.F.: Language resources for studying argument. In: Proceedings of the 6th Conference on Language Resources and Evaluation (LREC 2008), pp. 91–100 (2008)
2. Stab, C., Gurevych, I.: Annotating argument components and relations in persuasive essays. In: Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014), pp. 1501–1510 (2014)
3. Peldszus, A., Stede, M.: An annotated corpus of argumentative microtexts. In: First European Conference on Argumentation: Argumentation and Reasoned Action, Portugal, Lisbon (2015)
4. Walton, D., Reed, C., Macagno, F.: Argumentation Schemes. Cambridge University Press, Cambridge (2008)
5. Van Eemeren, F.H., Houtlosser, P., Snoeck Henkemans, F.: Argumentative Indicators in Discourse: A Pragma-Dialectical Study. Argumentation Library, vol. 12. Springer, Dordrecht (2007). https://doi.org/10.1007/978-1-4020-6244-5
6. Prasad, R., et al.: The Penn Discourse Treebank 2.0 Annotation Manual. Technical report 203, Institute for Research in Cognitive Science, University of Pennsylvania (2007)
7. Van Dejk, T.: Text and Context: Explorations in the Semantics and Pragmatics of Discourse. Longman, London (1977)

8. Mann, W.C., Thompson, S.A.: Rhetorical structure theory: toward a functional theory of text organization. Text **8**(3), 243–281 (1988)
9. Stede, M., Afantenos, S.D., Peldszus, A., Asher, N., Perret, J.: Parallel discourse annotations on a corpus of short texts. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC), Portoroz (2016)
10. Stab, C., Gurevych, I.: Identifying argumentative discourse structures in persuasive essays. In: Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, pp. 46–56 (2014)
11. Eckle-Kohler, J., Kluge, R., Gurevych, I.: On the role of discourse markers for discriminating claims and premises in argumentative discourse. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2236–2242 (2015)
12. Taboada, M., Das, D.: Annotation upon annotation: adding signalling information to a corpus of discourse relations. Dialogue Discourse **4**(2), 249–281 (2013)
13. Toldova, S., et al.: Rhetorical structure markers in Russian RST Treebank. In: Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms, pp. 29–33 (2017)
14. Toldova, S., Pisarevskaya, D., Vasilyeva, M., Kobozeva, M.: The cues for rhetorical relations in Russian: "Cause-Effect" relation in Russian Rhetorical Structure Treebank. In: Computational Linguistics and Intellectual Technologies, Papers from the Annual International Conference "Dialogue", pp. 747–761 (2018)
15. Rahwan, I., Reed, C.: The argument interchange format. In: Simari, G., Rahwan, I. (eds.) Argumentation in Artificial Intelligence. Springer, Boston (2009). https://doi.org/10.1007/978-0-387-98197-0_19
16. Gordon, T.F., Walton, D.: The Carneades argumentation framework–using presumptions and exceptions to model critical questions. In: 6th Computational Models of Natural Argument Workshop (CMNA), European Conference on Artificial Intelligence (ECAI), vol. 6, pp. 5–13 (2006)
17. Sidorova, E.A.: Multipurpose dictionary subsystem for extraction of subject lexicon. In: Proceedings of the International Conference "Computational Linguistics and Intellectual Technologies" (Dialogue-2008), pp. 475–481 (2008). (in Russian)
18. Kononenko, I., Ahmadeeva, I., Sidorova, E., Shestakov, V.: Problems of extracting terminological core of the subject domain from electronic encyclopedic dictionaries. Syst. Inform. **13**, 49–76 (2018). (in Russian)