Rhetorical relation markers in Russian RST Treebank

Svetlana Toldova (1), Dina Pisarevskaya (2), Margarita Ananyeva (3), Maria Kobozeva (3), Alexander Nasedkin (1), Sofia Nikiforova (1), Irina Pavlova (1), and Alexey Shelepov (1)
1 NRU Higher School of Economics, Moscow, Russia
2 Institute for Oriental Studies of the RAS, Moscow, Russia
3 Institute for Systems Analysis FRC CSC RAS
toldova@yandex.ru, dinabpr@gmail.com, ananyeva@isa.ru, kobozeva@isa.ru, kloudsnuff@gmail.com, son.nik@mail.ru, ispavlovais@gmail.com, alexshelepov1992@gmail.com

Abstract

The paper deals with the pilot version of the first RST discourse treebank for Russian. The project started in 2016. At present, the treebank consists of sixty news texts annotated for rhetorical relations according to RST scheme. However, this scheme was slightly modified in order to achieve higher inter-annotator agreement score. During the annotation procedure, we also registered the discourse connectives of different types and mapped them onto the corresponding rhetoric relations. In present paper, we discuss our experience of RST scheme adaptation for Russian news texts. Besides, we report on the distribution of the most frequent discourse connectives in our corpus.

1 Introduction

One of the focuses of the present NLP research is the text analysis on the discourse level. There is a big amount of NLP tasks, such as coreference resolution, text summarization, irony detection, question-answering systems etc., where the analysis of text needs to go beyond the boundaries of a single clause or even a sentence. For such tasks, the information on text cohesion, discourse structure and discourse relations is needed. In order to develop the modules dealing with discourse analysis, one needs a text corpus with discourse level annotation.

This paper describes the creation of the pilot version of the Discourse-annotated corpus for the Russian language, based on Rhetorical Structure Theory (RST) framework (Mann, Thompson, 1988). Corpus includes the texts taken from Russian freely available online resources and manually annotated for RST relations. It is designed for conducting the experiments on different machine-learning methods for discourse parsing. It also can be used for the investigation of discourse structure, relational and lexical cohesion and other discourse-based phenomena in Russian.

During the annotation procedure we single out different connectives (conjunctions, particles, some lexical and punctuation cues), associated with the corresponding discourse relation. These cues can serve as a seed set for automatic discourse connectives extraction.

Until now, the majority of theoretical works devoted to discourse relation for Russian were dealing primarily with the analysis of conjunction, parenthesis words and expressions functions. Our approach differs in that our goal was to find out what lexical items irrespective of their part of speech can signal the presence of a rhetorical relation. Thus, we take into consideration such lexical clues as nouns or verbs of speech etc. (e.g. *prichina* ‘the

Proceedings of the 6th Workshop Recent Advances in RST and Related Formalisms, pages 29–33,
Santiago de Compostela, Spain, September 4 2017.

course'). In present paper, we suggest quantitative analyses of these connectives.

2 Related works

There exist different approaches to discourse annotation principles. One of the approaches is based on the "linear" annotation. Thus, in Penn Discourse Treebank (PDTB) discourse relations are lexically anchored by discourse connectives. They are viewed as predicates that take abstract objects such as propositions, events and states as their arguments (PDTB (Prasad et al., 2007; Webber et al., 2016), TurkishDB (Zeyrek et al., 2013), etc.). In the Chinese Discourse TreeBank the punctuation marks also play role in the annotation (Zhou, Xue, 2015). Models based on cohesive relations are not tree-like, for instance, Discourse Graphbank (Wolf and Gibson, 2005). Another significant approach is the Rhetorical Structure Theory (RST) (Mann, Thompson, 1988). RST framework represents text as a hierarchy of elementary discourse units (EDUs) and describes relations between them and between bigger parts of text. Some EDUs carry more important information (nucleus) than others (satellite) do. There are two rhetorical relation types: nucleus-satellite (mononuclear) and multinuclear. While the first type connects a nucleus and a satellite, the latter includes EDUs that are equally important in the analyzed discourse. For the current research we chose RST to study cohesive markers and discourse cues taking into consideration 'trees' - discourse structure of texts.

There exist special lexicons or extensive descriptions of discourse connectives' (their types, positions, linking directions, ambiguous degrees, distribution of signalled relations) for particular languages: e.g. for English (Taboada M., Das D, 2013), for French (Roze C., Danlos L., Muller P. LEXCONN), for Chinese (Huang H. H. et al., 2014), etc. There are also comparatives studies of discourse connectives (e.g. English and French (Popescu-Belis A. et al, 2012), Spanish and Chinese (Cao S., da Cunha I., Bel N, 2016)).

As some discourse markers can indicate more than one discourse relation, another problem in this field is a lexical cue disambiguation (da Cunha I., 2013; Khazaei T. et al., 2015). General way of resolving this problem is extracting syntactic contexts for a particular cue in different discourse relation.

For automatic discourse parsing the most complicated task is to identify implicit discourse relations - those that do not involve any explicit discourse connectives. In (Rutherford A. et al., 2015) authors investigated the criteria for selecting the discourse connectives that can be omitted without changing the context.

M. Taboada and D. Das (Taboada M., Das D, 2013) suggest an exhaustive investigation of discourse relation clues. Besides traditionally discussed functional words, such as conjunctions, the list of connectives features is extended by means of semantic, syntactic, graphical and others types of features. As a result, authors show that the majority of relations are explicit rather than implicit, as it is usually postulated. Making a list of discourse relations clues for Russian, we take this approach into consideration.

3 Russian RST Bank

The current project started in 2016. We are planning to annotate texts (more than 100,000 tokens) of four genres and domains: science, popular science, news stories, and analytic journalism. The pilot project was aimed at working-out annotation rules and to achieve a reasonable score for inter-annotator agreement.

For annotation we use an open-source tool rstWeb [https://corpling.uis.georgetown.edu/rst-web/info/]. It has a number of advantages in comparison with other tools (UAM CorpusTool, RSTTool, GraphAnno): user-friendly interface, ability to work in the browser and to make changes to the code.

We start with the list of relations suggested in (Mann W., Tompson S., 1988). The instruction for annotators was based on the work by L.Carlson, D. Marcu, M. Okurowsky (Carlson et al., 2003). However, the initial list of relations was slightly modified. After modification, the resulting list consisted of 25 relations. During the further tagging procedure, special focus was on inter-annotator agreement (IAA). We have selected Krippendorff's unitized alpha as a statistic to measure IAA. It operates on the whole annotation spans instead of isolated tokens and it can be calculated for any number of annotators.

It turned out that annotators confuse Volitional and Non-Volitional relations, Antithesis and Con-

trast (same meaning, but Antithesis is mononuclear, Contrast - multinuclear), Cause and Effect (same meaning, but either cause or effect is nuclear). We decided to tack them, as well as two types of Attribution (a general and more specific one) and Interpretation with Evaluation (the only difference is in the degree of objectivity of author's evaluation). Besides, we took out Conclusion and Motivation, since they occur rarely and the first one can be considered a subtype of Restatement). Finally, we got 17 relations that were divided into four groups (fig. 1). These modifications have given a vast improvement of IAA. For three texts tagged by four people it stood at 0,27 - 0,49 before reduction of the relations tree and 0,69 - 0,77 after reduction. In order to accelerate the annotation process, the automatic text segmentation was

1. **Coherence**
   - 1.1. Background
   - 1.2. Elaboration
   - 1.3. Restatement
   - 1.4. Interpretation - Evaluation
   - 1.5. Preparation
   - 1.6. Solutionhood
2. **Casual-argumentative**
   - 2.1. Contrastive
     - 2.1.1. Concession
     - 2.1.2. Contrast
   - 2.2. Causal
     - 2.2.1. Purpose
     - 2.2.2. Evidence
     - 2.2.3. Cause-Effect
   - 2.3. Condition
3. **Structural**
   - 3.1. Sequence
   - 3.2. Joint
   - 3.3. Same-unit
   - 3.4. Comparison
4. **Attribution**
   - 4.1. Attribution

Figure 1. The list of rhetoric relations

applied. RusClaSp (http://gree-gorey.github.io/) package was taken as a basis and adapted to our task and corpus. In particular, we consider some explicit unambiguous markers and ignore parenthetic phrases. The human-annotator checks the result of automatic segmentation and builds a discourse tree of a text.

By now, we have annotated 73 texts, mostly news stories (each of them is 30 sentences in length on the average), they contain 44685 tokens. For each text we built one single tree where text spans are connected to other spans, nodes are connected to other nodes, and so on to the common vertex.

4 Rhetorical relations markers

In our current research, we investigate the interaction between discourse connectives and the discourse relations. As it has been already mentioned, we consider not only functional words to be rhetoric relation markers. The markers includes punctuation marks, prepositions, pronouns, speech verbs, etc.

While annotating the corpus, we register overt clues for corresponding relation types. The list of registered cases consists of 692 pairs "marker-relation" (with approximately 200-250 unique markers suggested by annotators). The variation in number of markers due to the fact that some the markers are constructions where one of their elements may vary. For instance, one of the patterns for ATTRIBUTION relation is a construction introducing "reported speech" consisting of a verb of speech plus, optionally, a conjunction *chto* 'what, that' (e.g. "said that" or "reported that" etc.). There is no enough data to decide whether to treat the elements of this construction as separate markers or not.

Markers, which appear in texts more frequently, may be ambiguous, i.e. same markers can signal several relations. There are 55 markers ordered by the raise of their frequencies (threshold >= 3 occurrences in the corpus) in fig. 2.

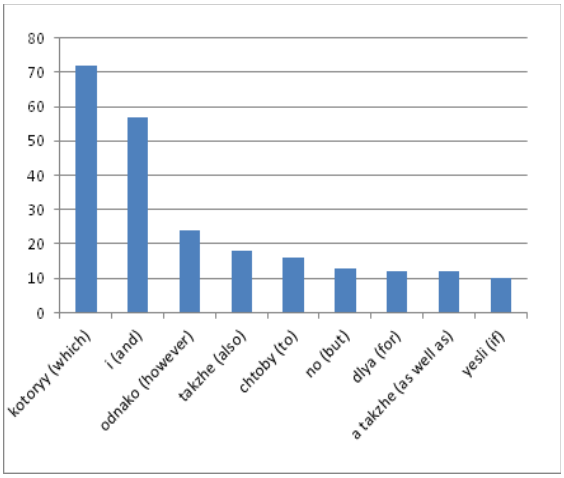

Figure 2. The frequency of top 55 markers

Among 14 most frequent markers (threshold >=9), most of them (with one exception of *v to vremya,, kak* 'at time, when') point directly on the definite relation type or close relation types. The table 1 presents the statistics for relations expressed overtly via markers. The most frequent marker for this relation is given.

Relation type	Freq	marker	translation
Elaboration	150	*kotoryj*	"which, that"
Joint	119	*i, takzhe*	and, as well
Attribution	118	*zajavil, soobschil*	report, announce etc.
Contrast	62	*Odnako, a, no*	However, but
Cause-Effect	47	*Poetomu, V+prichina*	so, accordingly, V+*cause*
Purpose	39	*Chtoby, dlya*	In order that, for
Interpretation-Evaluation	34	Nouns and verbs expressing opinion	
Background	31	No dominant marker	
Condition	27	*esli*	if

Table 1. Relations with their most frequent markers

As we can see from the table the most frequent relation in News texts are ELABORATION, JOINT and ATTRIBUTION. These texts are characterized by high proportion of symmetric relations and high quantity of special lexical expressions such as constructions with speech verbs and other types of mental predicates.

5 Discussion

The discourse markers analysis reveals some interesting evidence that deserves additional attention. Firstly, the news texts contain not many special subordinate conjunctions for reason, cause etc. The most frequent are such relations as JOINT, ELABORATION and ATTRIBUTION.

The punctuation marks in Russian such as hyphen can also signal some relations, namely, ELABORATION.

The JOINT relation is expressed not only via coordinative conjunction, but also via the conjunction *a* "but" traditionally treated as adversative.

The clause type for elaboration in Russian news texts is a relative clause (finite clause or participial clause). Thus, the marker for elaboration is the relative pronoun *kotoryj* 'which'.

The task to extract the ATTRIBUTION relation can be reformulated as the task to extract the markers of reported speech. Almost all the markers that the annotators single out for ATTRIBUTION are special constructions for reported speech introduction into discourse such as 'said that', 'according to X's opinion', 'As X's announced…'

There is a tendency in News texts to express cause-effect and some other relations via special lexemes denoting some mental operations (assessment, intentions etc.).

6 Conclusions

The aim of this paper was to introduce an ongoing project on a new RST TreeBank construction and to discuss our experience of adopting the RST scheme for rhetoric relations annotation for Russian. We also have provided a pilot research of different types of discourse clues. We are going to use some of these clues as a seed set for bootstrapping some other discourse markers and map them for specific rhetoric relations. The survey of different markers extracted by the annotators is helpful for feature extraction for developing a discourse parser for Russian based on machine learning.

Acknowledgements

This research was supported by a grant from Russian Foundation for Basic Research Fund (15-07-09306).

References

A. Popescu-Belis, T. Meyer, J. Liyanapathirana, B.Cartoni, and Zufferey S. 2012. Discourse-level annotation over Europarl for machine translation: Connectives and pronouns. *In Proceedings of the eighth international conference on Language Resources and Evaluation (LREC)*. № EPFL-CONF-192582: 2716-2720.

A. Rutherford and N. Xue. 2015. Improving the Inference of Implicit Discourse Relations via Classifying Explicit Discourse Connectives. *HLT-NAACL*: 799-808.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2016. A Discourse-Annotated Corpus of Conjoined VPs. *Proc. 10th Linguistics Annotation Workshop*, Berlin: 22–31.

C. Roze, L. Danlos, and P. Muller. 2012. LEXCONN: a French lexicon of discourse connectives. In Discours. Revue de linguistique, psycholinguistique et informatique. *A journal of linguistics, psycholinguistics and computational linguistics.* №. 10.

D. Zeyrek, I. Demirşahin, A.B. Sevdik Çallı, and R. Çakıcı. 2013. Turkish Discourse Bank: Porting a discourse annotation style to a morphologically rich language. *Dialogue and Discourse, 4(2)*: 174–184.

Florian Wolf and Edward Gibson. 2005. Representing discourse coherence: *A corpus-based study. In Computational Linguistics,* 31(2): 249–287.

H.-H. Huang, T.-W. Chang, H.-E. Chen, and H.-H. Chen. 2014. Interpretation of Chinese Discourse Connectives for Explicit Discourse Relation Recognition. *Proceedings of COLING* 2014: 632-643.

Iria da Cunha, Juan-Manuel Torres-Moreno, and Gerardo Sierra. 2011. On the development of the RST Spanish treebank. *Proceedings of the 5th Linguistic Annotation Workshop (LAW V)*: 1–10.

L. Carlson, D. Marcu, M.E. Okurowski. 2003. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory, Current directions in discourse and dialogue, Kluwer Academic Publishers: 85–112.

M. Taboada. 2013. Das D. Annotation upon Annotation: Adding Signalling Information to a Corpus of Discourse Relations. In D&D. Vol.. 4. №. 2: 249-281.

N. van der Vliet, I. Berzlanovich, G. Bouma, M. Egg, and G. Redeker. 2011. Building a Discourse-Annotated Dutch Text Corpus. *Proceedings of the Workshop "Beyond Semantics: Corpus-based Investigations of Pragmatic and Discourse Phenomena"*, Goettingen, Germany, 23–25 February 2011: 157–171.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie Webber. 2007. The Penn Discourse Treebank 2.0 Annotation Manual. Technical Report 203, Institute for Research in Cognitive Science, University of Pennsylvania.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn discourse treebank 2.0. *Proceedings of the 6th International Conference on Language Resources and Evaluation*: 2961–2968.

S. Cao, I. da Cunha, and N. Bel. 2016. An analysis of the Concession relation based on the discourse marker aunque in a Spanish-Chinese parallel corpus. *Procesamiento del Lenguaje Natural.* Vol. 56: 81-88.

S. Joty, G. Carenini, and R. T. Ng. 2015. CODRA: A Novel Discriminative Framework for Rhetorical Analysis. In Computational Linguistics 41, 3: 385-435.

S.Y. Cao, I. da Cunha, and M. Iruskieta. 2016. Elaboration of a Spanish-Chinese parallel corpus with translation and language learning purposes, *34th International Conference of the Spanish Society for Applied Linguistics (AESLA)*, to appear.

I. da Cunha. 2013. A Symbolic Corpus-based Approach to Detect and Solve the Ambiguity of Discourse Markers, *Research in Computing Science* 70: 93-104.

T. Khazaei, L. Xiao, and R.E. Mercer. 2015. Identification and Disambiguation of Lexical Cues of Rhetorical Relations across Different Text Genres. Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics (LSDSem): 54-63.

W.C. Mann and S.A. Thompson. 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization, *Text 8*, 3, 1988: 243–281.

Y. Zhou and N. Xue. 2015. The Chinese Discourse TreeBank: A Chinese corpus annotated with discourse relations. *Language Resources and Evaluation*: 397–431.