HIGHER SCHOOL OF ECONOMICS

NATIONAL RESEARCH UNIVERSITY

Federal Research Centre
Computer Science and Control
Russian Academy of Sciences

Towards Building a Discourse-annotated Corpus of Russian

Speakers: Dina Pisarevskaya and Margarita Ananyeva

Background: Discourse analysis

Can be useful in **Natural Language Processing tasks**:

- machine translation evaluation,
- sentiment analysis,
- information retrieval,
- information extraction,
- text summarization,
- anaphora resolution,
- question-answering systems,
- text classification.

Discourse parsers for English:
RASTA [Corston-Oliver, Corston-Oliver, 1998],
SPADE [Soricut, Marcu, 2003],
HILDA [Hernault et al., 2010],
CODRA [Joty et al., 2015].
Two parsers [Surdeanu et al., 2015]

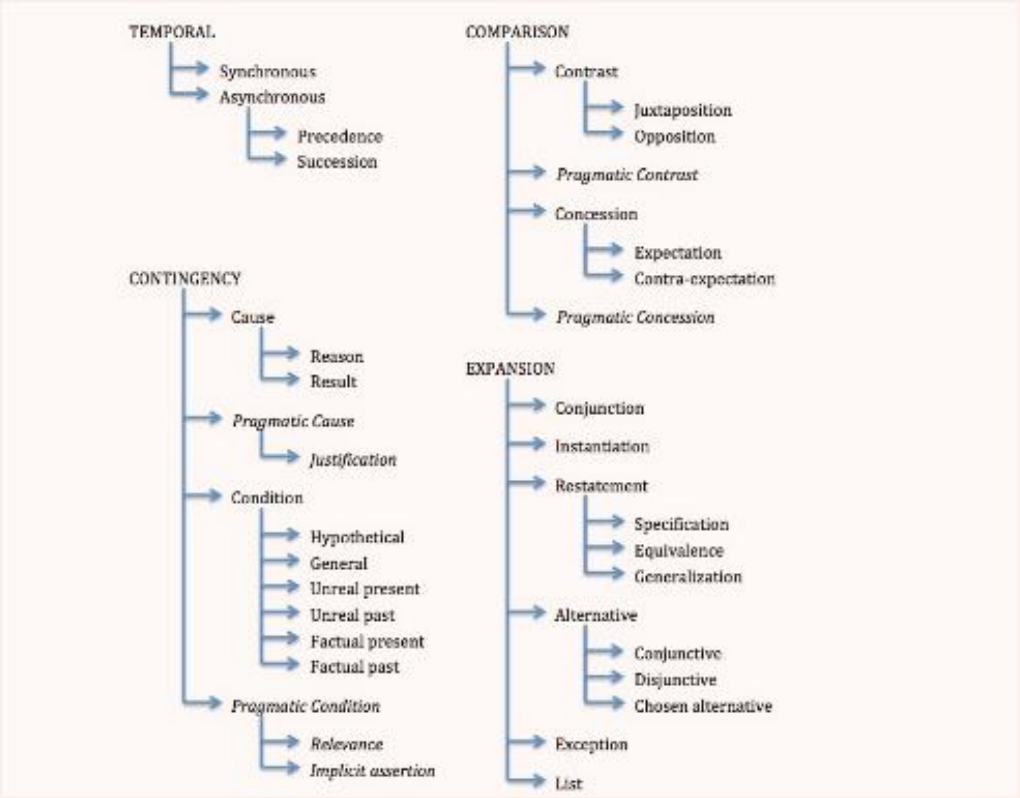
Discourse analysis approaches

- PDTB: Connective-led annotation (Penn Discourse Treebank) or Punctuation-led annotation (Chinese Discourse TreeBank). Example: PDTB (2008): 43 relations;
- Cohesive relations (Discourse Graphbank);
- Segment-led annotation (Rhetorical Structure Theory: a non-projective tree). Example: RST-DT (2003): 78 relations.

Penn Discourse Treebank

- Low-level relations (within/between adjacent sentences);
- Focus on discourse connectives;
- Relations have two (and only two) arguments.
- 3 levels of relation labels: class (4 major semantic classes), type (emphasizes the semantics of the class levels), subtype (emphasizes semantic contribution of each argument)
- When an annotator is uncertain of subtype, it is possible to choose higher level (type), it is good for inter-annotator agreement.

Penn Discourse Treebank: Relations



Penn Discourse Treebank: Corpora

Original corpus:

English: Penn Discourse Treebank (newspaper texts, million words).

Related corpora:

Chinese Discourse Treebank (newspaper texts, 70,000 words);

Czech: Prague Discourse Treebank (newspaper texts, 50,000 sentences);

6 languages: Eng, Tur, Deu, Por, Pol, Rus: TED-MDB (TED talks, work in progress);

Hindi: Discourse Relation Bank (newspaper texts, 400,000 words);

Arabic: Leeds Arabic DTB (newspaper texts, 166,000 words);

Turkish: METU-TDB Corpus (different genres, 500,000 words)

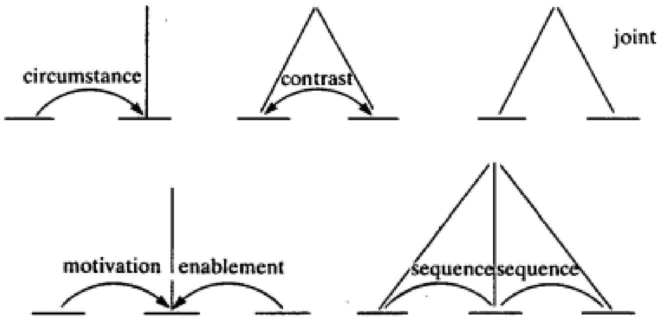
Discourse analysis approaches

- PDTB: Connective-led annotation (Penn Discourse Treebank) or Punctuation-led annotation (Chinese Discourse TreeBank). Example: PDTB (2008): 43 relations;
- Cohesive relations (Discourse Graphbank);
- Segment-led annotation (Rhetorical Structure Theory: a non-projective tree). No strong focus on connectives like in PDTB. Example: RST-DT (2003): 78 relations.

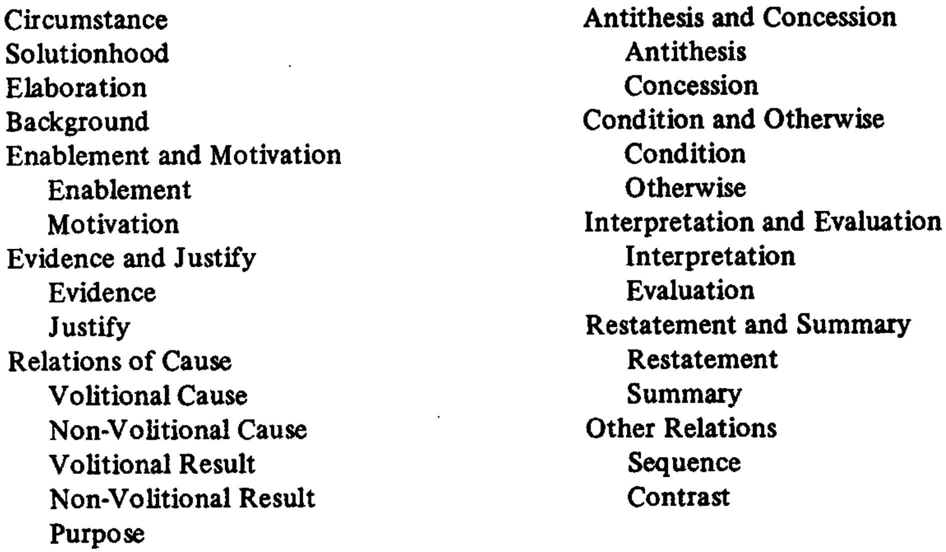
Rhetorical Structure Theory

[Mann, Thompson, 1988]

Examples of schema types



"Classic" relations set



- Circumstance
- Solutionhood
- Elaboration
- Background
- Enablement and Motivation
  - Enablement
  - Motivation
- Evidence and Justify
  - Evidence
  - Justify
- Relations of Cause
  - Volitional Cause
  - Non-Volitional Cause
  - Volitional Result
  - Non-Volitional Result
  - Purpose
- Antithesis and Concession
  - Antithesis
  - Concession
- Condition and Otherwise
  - Condition
  - Otherwise
- Interpretation and Evaluation
  - Interpretation
  - Evaluation
- Restatement and Summary
  - Restatement
  - Summary
- Other Relations
  - Sequence
  - Contrast

RST-corpora for different languages

- **English:** RST Discourse Treebank [Carlson et al., 2003], 385 newspaper articles, 176 383 tokens
- **German:** Potsdam Commentary Corpus [Stede, Neumann, 2014], 2 900 sentences from 175 newspaper articles, 32 000 tokens
- **Portuguese:** CorpusTCC [Pardo et al., 2004], 1 350 sentences from 100 scientific texts, 53 000 tokens
- **Portuguese:** Rhetalho [Pardo et al., 2004], 50 texts (30 from scientific papers and 20 from newspaper), approximately 5 000 tokens
- **Spanish:** RST Spanish Treebank [da Cunha et al., 2011], 2 256 sentences from 267 documents of several genres
- **Japanese:** [Kawahara et al., 2014], 30 000 sentences from 10 000 documents from the web, variety of domains

Discourse-annotated corpus of Russian

Texts of 4 genres:

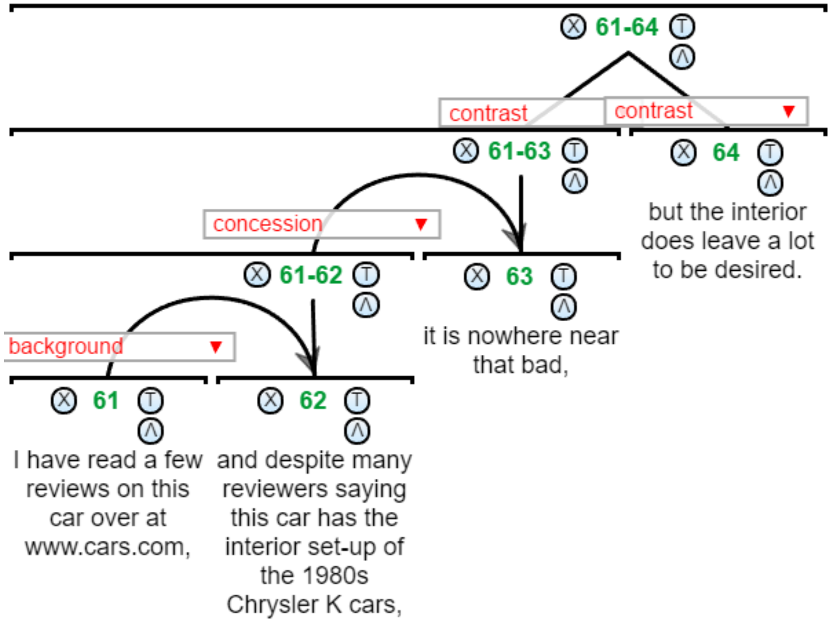
- science;
- popular science;
- news stories;
- analytic journalism.

T**he project:**

- 3 years;
- > 100 texts;
- > 100 000 tokens.

Annotation Tool

Open-source annotation tool rstWeb [https://corpling.uis.georgetown.edu/rstweb/info/]



I have read a few reviews on this car over at www.cars.com,

and despite many reviewers saying this car has the interior set-up of the 1980s Chrysler K cars,

it is nowhere near that bad,

but the interior does leave a lot to be desired.

Background

D. Pisarevskaya,
“Rhetorical Structure Theory as a Feature for Deception Detection in News Reports in the Russian Language”
- Master thesis in Higher School of Economics, Computational Linguistics (the results were presented on 1st June).

Background (2)

The Laboratory for Computer Linguistics and Intelligent Information Processing (Institute for Systems Analysis FRC CSC RAS).

- manual (21 relations);
- 10 texts (1200 units and 1484 relations) from SynTagRus;
- discourse markers.

Kobozeva M.
“Developing the corpus of Russian texts with markup based on the Rhetorical Structure Theory”
- Master thesis in Russian State University for Humanities, Computational Linguistics
Ananyeva M. I., Kobozeva M. B. (2016), Developing the corpus of Russian texts with markup based on the Rhetorical Structure Theory, “Dialogue 2016”

Current research

New corpus - 60 news stories have already been annotated.

User manual has been updated.

Segmentation of Russian texts into clauses: http://gree-gorey.github.io/

Inter-annotator agreement

- Accuracy
- Cohen’s kappa coefficient [Cohen, J., 1960; Cohen, J., 1968]
  - Scott’s Pi [Scott, W. A., 1955]
- Token-based Fleiss’ kappa [Fleiss, J. L., 1971]
- Krippendorff’s unitized alpha [Krippendorff K., 2007]

Relations

Mononuclear	**Multinuclear**
1. Background 2. Volitional and Non-Volitional Cause 3. Evidence 4. Volitional and Non-Volitional Effect 5. Condition 6. Purpose 7. Concession 8. Preparation 9. Conclusion 10.Elaboration 11.Antithesis 12.Solutionhood 13.Motivation 14.Evaluation 15.Interpretation 16.Attribution1 and Attribution2	1. Contrast 2. Restatement 3. Sequence 4. Joint 5. Comparison 6. Same-unit

Evolution of relations

Volitional Cause + Non-volitional Cause = Cause

→ **Cause + Effect**

Volitional Effect + Non-volitional Effect = Effect

Interpretation + Evaluation

Attribution1 + Attribution2 = Attribution

~~**Antithesis**~~

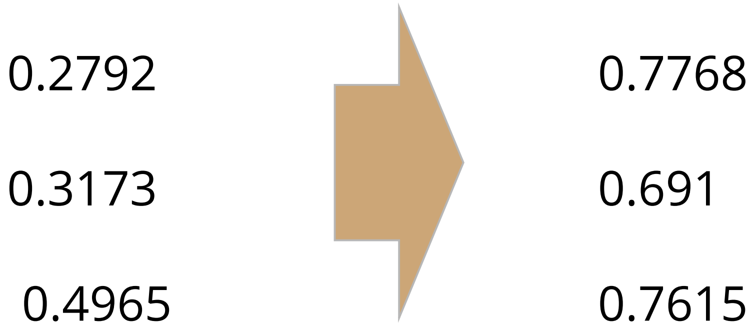
~~**Conclusion**~~

~~**Motivation**~~

New RST relations tree

Coherence	**Casual-argumentative**	**Structural**	**Attribution**
Background	Contrastive	Sequence	Attribution
Elaboration	Concession	Joint	
Restatement	Contrast	Same-unit	
Interpretation - Evaluation	Causal	Comparison	
Preparation	Purpose		
Solutionhood	Evidence		
	Cause-Effect		
	Condition		

Inter-annotator agreement



0.2792 → 0.7768

0.3173 → 0.691

0.4965 → 0.7615

The code used for IAA calculation can be accessed via GitHub [https://github.com/nasedkinav/rst_corpus_rus/blob/master/krippendorffs_alpha.py].

Future work

User-friendly interface: visualisation, search and statistics, file upload mechanism.

Analysis of “marker potential”.

Discourse parser.

References

Cao S. Y., da Cunha I., Iruskieta M. (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.

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

Corston-Oliver S., Corston-Oliver S. H. (1998), Beyond string matching and cue phrases: Improving efficiency and coverage in discourse analysis. In The AAAI Spring Symposium on Intelligent Text Summarization, pp. 9–15.

da Cunha I., Torres-Moreno J.-M., Sierra G. (2011), On the development of the RST Spanish treebank. In Proceedings of the 5th Linguistic Annotation Workshop (LAW V), pp. 1–10.

Hernault H., Prendinger H., duVerle D., Ishizuka M. (2010), HILDA: A discourse parser using support vector machine classification. In Dialogue & Discourse, 1(3), pp. 1–33.

Iruskieta M., Aranzabe M. J., Díaz de Ilarraza A., Gonzalez I., Lersundi M., Lopez de la Calle O. (2013), The RST Basque TreeBank: an online search interface to check rhetorical relations, IV Workshop RST and Discourse Studies. Fortaleza, Brasil, Outubro 21-23, pp. 40-49.

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

References

Kawahara D., Machida Y., Shibata T., Kurohashi S., Kobayashi H., Sassano M. (2014), Rapid Development of a Corpus with Discourse Annotations using Two-stage Crowdsourcing. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 269–278.

Mann W. C., Thompson S. A. (1988), Rhetorical Structure Theory: Toward a Functional Theory of Text Organization , Text 8, 3, 1988, pp. 243-281.

Pardo T. A. S., Nunes M. G. V., Rino L. H. M. (2004), Dizer: An automatic discourse analyzer for brazilian portuguese, Brazilian Symposium on Artificial Intelligence, Springer Berlin Heidelberg, pp. 224-234.

Soricut R., Marcu D. (2003), Sentence Level Discourse Parsing Using Syntactic and Lexical Information. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, NAACL’03, pp. 149–156..

Stede M., Neumann A. (2014), Potsdam Commentary Corpus 2.0: Annotation for Discourse Research. Proc. of LREC, Reykjavik.

Surdeanu M., Hicks T., Valenzuela-Esćarcega M.A. (2015), Two Practical Rhetorical Structure Theory Parsers, Proceedings of NAACL-HLT 2015, pp. 1–5.

References

Van der Vliet N., Berzlanovich I., Bouma G., Egg M., Redeker G. (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, pp. 157-171.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and psychological measurement, 20(1), 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological bulletin, 70(4), 213.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. Psychological bulletin, 76(5), 378.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. Public opinion quarterly, 321-325.

Krippendorff, K. (2007). Computing Krippendorff's alpha reliability. Departmental papers (ASC), 43.

Thank you for attention

Dina Pisarevskaya

dinabpr@gmail.com

https://github.com/nasedkinav/rst_corpus_rus

HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY

Federal Research Centre
Computer Science
and Control
Russian Academy of Sciences