Annotating coherence relations in corpora of language use

Merel C.J. Scholman, Ted J.M. Sanders
Utrecht institute of Linguistics OTS
Trans 10, 3512 JK Utrecht, the Netherlands
E-mail: m.c.j.scholman@uu.nl, t.j.m.sanders@uu.nl

Keywords: Corpora, discourse annotation, text analysis

1. Corpora and their influence on the study of language

The advent of linguistic corpora has had a large impact on the field of linguistics and its research community. By gathering and annotating large-scale collections of text, researchers have gained new possibilities for analyzing language. The focus area of corpora has mainly been on lexical, syntactic and semantic characteristics of language. Existing corpora often lack annotations on the discourse level (Carlson, Marcu & Okurowski, 2001). However, the notion of "discourse", and more specifically the coherence relations between parts of discourse such as *cause-consequence* and *claim-argument*, has become increasingly important in linguistics over the years. In the last decennium there has been an international tendency to create discourse annotated corpora. Leading examples are the Penn Discourse Treebank (Prasad et al., 2008) and the RST Treebank (Carlson et al., 2001). While discourse annotation guidelines generally agree on the idea of coherence relations, a uniform standard for discourse annotation is not yet available. Moreover, the agreement between analysts on certain values is often problematic.

In the context of a CLARIN-NL project, Sanders, Broeder and Vis (2012) developed a systematic method to annotate coherence relations, by adjusting an existing theory on categories of coherence relations (Sanders, Spooren & Noordman, 1992) to allow for a step-wise annotation process. This annotation scheme has been used to annotate a discourse corpus for Dutch, named DiscAn, which can be accessed in the MPI Archives (http://www.mpi.nl/resources/data), where it can be browsed. The DiscAn-corpus currently contains approximately 1500 fragments of written, spoken and chat text in Dutch.

In this contribution we first introduce the cognitive approach to coherence relations (CCR) used to annotate the DiscAn corpus, after which we present two annotation experiments designed to investigate the usability and reliability of this approach for discourse annotation.

2. Categories of coherence relations: a cognitive approach

Coherence relations are considered to be cognitive elements of the discourse representation (Sanders et al., 1992). In order to identify and describe coherence relations, Sanders, Spooren and Noordman (1992) distinguish four cognitive categories which they claim to be relevant for every coherence relation. These are: basic operation (relations are causal or additive), source of coherence (objective or subjective), order of the segments (basic or non-basic order), and polarity (positive or negative). These four categories allow for a systematic, step-wise analytical process and can be visualized in a flowchart. Other annotation schemes often lack a theoretical framework. Consequently, the schemes are not organized systematically and contain counter-intuitive aspects. For example, in the PDTB, contrastive relations expressed by the connective *but* can end up in different classes of relations. We expect this to be confusing to annotators, which is why they need large manuals and intensive training. However, a more systematically organized approach might make it possible to use non-expert annotators and fewer instructions. Thus, the systematical, step-wise approach of CCR is expected to be beneficial to annotators.The present study therefore aims to investigate to what extent the cognitive approach to coherence relations can be used reliably by non-trained, non-expert annotators in discourse annotation.

3. Investigating the usability of the cognitive categories in discourse annotation

Twenty non-trained, non-expert undergraduate students analyzed a sample corpus of 60 fragments which were taken from the DiscAn-corpus. Subjects were asked to annotate the sample corpus using a manual and an instruction. Two versions of the instruction were created. Version 1, the implicit instruction, relies only on the annotator's knowledge of the categories. Version 2, the explicit instruction, relies on this knowledge, as well as on paraphrase and substitution tests. These types of tests make use of connective properties and paraphrases of the two segments of a coherence relation (Knott & Dale, 1994). Paraphrase and substitution tests facilitate the annotation process and are therefore expected to lead to more agreement between annotators than the implicit instruction. The manual and instruction were presented on paper, whereas the corpus was presented in an excel-file on a computer. Subjects were instructed to fill in their annotations in this excel-file.

3.1 Results

Kappa statistics were used to calculate the agreement with annotators and the original annotations, which were done by an expert annotators for the DiscAn-corpus. Table 1 shows the agreement statistics.

Category	Overall	Implicit instruction	Explicit instruction
Polarity	.81	.78	.85
Basic operation	.44	.49	.40
S. of coherence	.25	.21	.29
Order	.66	.69	.64

Table 1: Agreement with original annotations in kappa statistics

Category	Overall	Implicit instruction	Explicit instruction
Polarity	.86	.81	.89
Basic operation	.49	.45	.53
S. of coherence	.31	.34	.28
Order	.61	.54	.65

Table 2: Agreement with original annotations in kappa statistics

Table 1 shows that agreement with original scores on *polarity* is almost perfect (κ = .81) and agreement on *order* is substantial (κ = .66). Agreement on *basic operation* is moderate (κ = .44), and agreement on *source of coherence* is fair (κ = .25). Further analyses only revealed a significant difference between the instructions for the category basic operation (t = 3.17; df = 1181,59; p = 0.002): participants using the implicit instruction show higher agreement than participants using the explicit instruction for the category basic operation.

The results indicate that the cognitive categories method can partly be applied reliably by non-trained, non-expert annotators. Tentative conclusions can be drawn for the categories *polarity* and *order*, but agreement was not adequate for *basic operation* and *source of coherence*. The results also indicated that the explicit instruction did not benefit the agreement scores. One possible explanation for this is that the annotators in the explicit condition did not use the instruction at all times. As the instructions were presented on paper and the fragments on the computer, it was possible for annotators to skip steps in the instruction. To further investigate the influence of the type of instruction, a second, paper-and-pencil experiment was conducted.

4. Further investigation of the influence of instructions on annotation reliability

Forty non-trained, non-expert bachelor students took part in the second experiment. It was ensured that annotators could not skip steps in the instructions by converting both instructions into forms in which annotators had to tick the correct value. Thus, each fragment was presented on paper and followed by the instruction. For each step of the instruction, annotators had to tick their answer and thus provide their annotations on paper. No other changes were made to the manual, instructions and procedure.

4.1 Results

Table 2 reports agreement between annotators and the original annotations, done by an expert annotators. Agreement with original scores on *polarity* is almost perfect (κ = .86) and agreement on *order* is substantial (κ = .61). Agreement on *basic operation* is moderate (κ = .49), and agreement on *source of coherence* is fair (κ = .31). These results are similar to those of the first experiment. Further analyses revealed differences between the first and the second experiment: participants using the explicit instruction showed significantly more agreement than those using the implicit instruction on the categories polarity (t = -2.19; df =1356.83; p = .03), basic operation (t = -3.33; df = 1427.10; p = .001) and order (t = 3.32; df = 1418.45; p = .001). There was no significant difference in agreement with original annotations for the category source of coherence (t = 1.10; df = 1425.57; p = .27).

Results from this second experiment thus show that the type of instruction does influence how reliably annotators can analyze text: participants showed more agreement when annotating with the explicit instruction than with the implicit instruction As in Experiment 1, the category *polarity* yielded the highest agreement and *source of coherence* yielded the lowest agreement.

5. Conclusion

We have shown that the cognitive approach to coherence relations allows for a systematical, step-wise annotation process with which naïve annotators can yield considerable amounts of agreement. Analyzing coherence relations is a difficult task, even with extensive training and experience. For example, a study investigating the reliability of RST, using expert annotators, showed a kappa ranging from .6 to 1.0 (Carlson et al., 2001) and a study investigating the PDTB annotation scheme resulted in percentages of agreement ranging from 59.6 to 95.7 (Miltsakaki et al., 2004). In our study, non-trained, non-expert annotators using the cognitive approach to coherence relations manage to reach fair to almost perfect agreement, with percentages of agreement ranging from 56.7 (for *source of coherence*) to 95.5 (for *polarity*). Given that these annotators only received little instructions, the amounts of agreement they show are promising.

The current study also shows that an explicit instruction which includes substitution and paraphrase tests benefits annotator agreement. This might be generalized to other annotation schemes too. Further investigation could show whether substitution and paraphrase tests are beneficial for discourse annotation in general.

The results can be taken as a clue to the viability of this approach. It is likely that the approach described in this article will yield more agreement with expert annotators, who usually annotate discourse for corpora. But even non-expert annotators are likely to reach even higher agreement if they receive a short training phase or a slighty more detailed manual.

6. Acknowledgements

The authors would like to thank CLARIN-NL for their grant, which enabled them to pursue this work.

7. References

Carlson, L.; Marcu, D. and Okurowski, M.E. (2001). Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. *Proceedings of the 2nd SIGDIAL Workshop on Discourse and Dialogue*, Eurospeech 2001, Denmark.

Knott, A. and Dale, R. (1994). Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes*, 18: 35–62.

Miltsakaki, E., Joshi, A., Prasad, R. & Webber, B. (2004). Annotating Discourse Connectives and Their Arguments. In: *Proceedings of the Frontiers in Corpus Annotation 2004 NAACL/HLT Conference Workshop, Boston.*

Prasad, R.; Dinesh, N.; Lee, A.; Miltsakaki, E.; Robaldo, L.; Joshi, A. and Webber, B. (2008). The Penn Discourse Treebank 2.0. *Proceedings of the 6th International Conference of Language Resources and Evaluation (LREC 2008)*, Marrakech.

Sanders, T.J.M.; Spooren, W.P.M.S. and Noordman, L.G.M. (1992). Toward a Taxonomy of Coherence Relations. *Discourse Processes*, 15: 1–35.

Sanders, T.J.M.; Vis, K. and Broeder, D. (2012). *Project notes of CLARIN project DiscAn: Towards a Discourse Annotation system for Dutch language corpora.* Project notes. Utrecht University: Utrecht.