Automatic detection of causal relations in German multilogs

Tina Bögel **Annette Hautli-Janisz** **Sebastian Sulger** **Miriam Butt**

Department of Linguistics
University of Konstanz
firstname.lastname@uni-konstanz.de

Abstract

This paper introduces a linguistically-motivated, rule-based annotation system for causal discourse relations in transcripts of spoken multilogs in German. The overall aim is an automatic means of determining the degree of justification provided by a speaker in the delivery of an argument in a multiparty discussion. The system comprises of two parts: A disambiguation module which differentiates causal connectors from their other senses, and a discourse relation annotation system which marks the spans of text that constitute the reason and the result/conclusion expressed by the causal relation. The system is evaluated against a gold standard of German transcribed spoken dialogue. The results show that our system performs reliably well with respect to both tasks.

1 Introduction

In general, causality refers to the way of knowing whether one state of affairs is causally related to another.[1] Within linguistics, causality has long been established as a central phenomenon for investigation. In this paper, we look at causality from the perspective of a research question from political science, where the notion is particularly important when it comes to determining (a.o.) the *deliberative* quality of a discussion. The notion of deliberation is originally due to Habermas (1981), who assumes that within a deliberative democracy, stakeholders participating in a multilog, i.e. a multi-party conversation, justify their positions truthfully, rationally and respectfully and eventually defer to the better argument. Within political science, the question arises whether actual multilogs conducted in the process of a democratic decision making indeed follow this ideal and whether/how one can use automatic means to analyze the degree of deliberativity of a multilog (Dryzek (1990; 2000), Bohman (1996), Gutmann and Thompson (1996), Holzinger and Landwehr (2010)). The disambiguation of causal discourse markers and the determination of the relations they entail is a crucial aspect of measuring the deliberative quality of a multilog. In this paper, we develop a system that is designed to perform this task.

We describe a linguistically motivated, rule-based annotation system for German which disambiguates the multiple usages of causal discourse connectors in the language and reliably annotates the reason and result/conclusion relations that the connectors introduce. The paper proceeds as follows: Section 2 briefly reviews related work on the automatic extraction and annotation of causal relations, followed by a set of examples that illustrate some of the linguistic patterns in German (Section 3). We then introduce our rule-based annotation system (Section 4) and evaluate it against a hand-crafted gold standard in Section 5, where we also present the results from the same annotation task performed by a group of human annotators. In Section 6, we provide an in-depth system error analysis. Section 7 concludes the paper.

2 Related work

The automatic detection and annotation of causality in language has been approached from various angles, for example by providing gold-standard, (manually) annotated resources such as the Penn Discourse Treebank for English (Prasad et al., 2008), which was used, e.g., in the disambiguation of English connectives by Pitler and Nenkova (2009), the Potsdam Commentary Corpus for German (Stede, 2004) and the discourse annotation layer of Tüba-D/Z, a corpus of written German text (Versley and Gastel, 2012). Training auto-

[1]This work is part of the BMBF funded eHumanities project *VisArgue*, an interdisciplinary cooperation between political science, computer science and linguistics.

Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language, pages 20–27,
Gothenburg, Sweden, April 26, 2014.

matic systems that learn patterns of causality (Do et al., 2011; Mulkar-Mehta et al., 2011b, inter alia) is a crucial factor in measuring discourse coherence (Sanders, 2005), and is beneficial in approaches to question-answering (Girju, 2003; Prasad and Joshi, 2008).

With respect to automatically detecting causal relations in German, Versley (2010) uses English training data from the Penn Discourse Treebank in order to train an English annotation model. These English annotations can be projected to German in an English-German parallel corpus and on the basis of this a classifier of German discourse relations is trained. However, as previous studies have shown (Mulkar-Mehta et al., 2011a, inter alia), the reliability of detecting causal relations with automatic means differs highly between different genres. Our data consist of transcriptions of originally spoken multilogs and this type of data differs substantially from newspaper or other written texts.

Regarding the disambiguation of German connectives, Schneider and Stede (2012) carried out a corpus study of 42 German discourse connectives which are listed by Dipper and Stede (2006) as exhibiting a certain degree of ambiguity. Their results indicate that for a majority of ambiguous connectives, plain POS tagging is not reliable enough, and even contextual POS patterns are not sufficient in all cases. This is the same conclusion drawn by Dipper and Stede (2006), who also state that off-the-shelf POS taggers are too unreliable for the task. They instead suggest a mapping approach for 9 out of the 42 connectives and show that this assists considerably with disambiguation. As this also tallies with our experiments with POS taggers, we decided to implement a rule-based disambiguation module. This module takes into account contextual patterns and features of spoken communication and reliably detects causal connectors as well as the reason and result/conclusion discourse relations expressed in the connected clauses.

3 Linguistic phenomenon

In general, causality can hold between single concepts, e.g. between 'smoke' and 'fire', or between larger phrases. The phrases can be put into a causal relation via overt discourse connectors like 'because' or 'as', whereas other phrases encode causality implicitly by taking into account world knowledge about the connected events. In this paper, we restrict ourselves to the analysis of explicit discourse markers; in particular we investigate the eight most frequent German causal connectors, listed in Table 1. The *markers of reason* on the left head a subordinate clause that describes the cause of an effect stated in the matrix clause (or in the previous sentence(s)). The *markers of result/conclusion* on the other hand introduce a clause that describes the overall effect of a cause contained in the preceding clause/sentence(s). In the genre of argumentation that we are working with, the "results" tend to be logical conclusions that the speaker sees as following irrevocably from the cause presented in the argument.

Reason 'because of'	**Result** 'thus'
da	daher
weil	darum
denn	deshalb
zumal	deswegen

Table 1: German causal discourse connectors

The sentences in (1) and (2) provide examples of the phenomenon of explicit causal markers in German in our multilogs. Note that all of the causal markers in Table 1 connect a result/conclusion with a cause/reason. The difference lies in which of these relations is expressed in the clause headed by the causal connector.

The constructions in (1) and (2) exemplify this.[2] In (1), *da* 'since' introduces the reason for the conclusion in the matrix clause, i.e., the reason for the travel times being irrelevant is that they are not carried out as specified. In (2), *daher* 'thus' heads the conclusion of the reason which is provided in the matrix clause: Because the speaker has never stated a fact, the accusation of the interlocutor is not correct.

There are several challenges in the automatic annotation of these relations. First, some of the connectors can be ambiguous. In our case, four out of the eight causal discourse connectors in Table 1 are ambiguous (*da, denn, daher* and *darum*) and have, in addition to their causal meaning, temporal, locational or other usages. In example (3), *denn* is used as a particle signaling disbelief, while *daher* is used as a locational verb particle, having, together with the verb 'to come', the interpretation

[2]These examples are taken from the Stuttgart 21 arbitration process, see section 5.1 for more information.

(1) Diese Fahrzeiten sind irrelevant, *da* sie so nicht gefahren werden.
Art.Dem travel time.Pl be.3.Pl irrelevant because they like not drive.Perf.Part be.Fut.3.Pl

Result/Conclusion **Reason**

'These travel times are irrelevant, because they are not executed as specified.'

(2) Das habe ich nicht gesagt, *daher* ist Ihr Vorwurf nicht richtig
Pron have.Pres.1.Sg I not say.Past.Part thus be.3.Sg you.Sg.Pol/Pl accusation not correct

Reason **Result/Conclusion**

'I did not say that, therefore your accusation is not correct.'

(3) Wie kommen Sie *denn daher*?
how come.Inf you.Sg.Pol then VPart
'What is your problem anyway?' (lit. 'In what manner are you coming here?')

(4) *Da* bin ich mir nicht sicher.
there be.Pres.1.Sg I I.Dat not sure
'I'm not sure about that.'

(5) Das kommt *daher*, dass keiner etwas sagt.
Pron come.Pres.3.Sg thus that nobody something say.Pres.3.Sg

Result/Conclusion **Reason**

'This is because nobody says anything.'

of 'coming from somewhere to where the speaker is' (literally and metaphorically). In a second example in (4), *da* is used as the pronominal 'there'.

Second, some of the causal connectors do not always work the same way. In (5), the result/conclusion connector *daher* does not head an embedded clause, rather it is part of the matrix clause. In this case, the embedded clause expresses the reason rather than the result/conclusion. A third challenge is the span of the respective reason and result. While there are some indications as to how to define the stretch of these spans, there are some difficult challenges, further discussed in the error analysis in Section 6.

In the following, we present the rule-based annotation system, which deals with the identification of phrases expressing the result and reason, along the lines illustrated in (1) and (2), as well as with the disambiguation of causal connectors.

4 Rule-based annotation system

The automatic annotation system that we introduce is based on a linguistically informed, handcrafted set of rules that deals with the disambiguation of causal markers and the identification of causal relations in text. As a first step, we divide all of the utterances into smaller units of text in order to be able to work with a more fine-grained structure of the discourse. Following the literature, we call these discourse units. Although there is no consensus in the literature on what exactly a discourse unit consists of, it is generally assumed that each discourse unit describes a single event (Polanyi et al., 2004). Following Marcu (2000), we term these *elementary discourse units* (EDUs) and approximate the assumption made by Polanyi et al. (2004) by inserting a boundary at every punctuation mark and every clausal connector (conjunctions, complementizers). Sentence boundaries are additionally marked.

The annotation of discourse information is performed at the level of EDUs. There are sometimes instances in which a given relation such as "reason" spans multiple EDUs. In these cases, each of the EDUs involved is marked/annotated individually with the appropriate relation.

In the following, we briefly lay out the two elements of the annotation system, namely the disambiguation module and the system for identifying the causal relations.

4.1 Disambiguation

As shown in the examples above, markers like *da, denn, darum* and *daher* 'because/thus' have a number of different senses. The results presented in Dipper and Stede (2006) indicate that POS tagging alone does not help in disambiguating the causal usages from the other functions, particularly not for our data type, which includes much noise and exceptional constructions that are not present in written corpora. As a consequence, we propose a set of rules built on heuristics, which take into account a number of factors in the clause in order to disambiguate the connector. To illustrate the underlying procedure, (6) schematizes part of the disambiguation rule for the German causal connector *da* 'since'.

(6) IF *da* is not followed directly by a verb AND no other particle or connector precedes *da* AND
da is not late in the EDU THEN
da is a causal connector.

In total, the system comprises of 37 rules that disambiguate the causal connectors shown in Table 1. The evaluation in Section 5 shows that the system performs well overall.[3]

4.2 Relation identification

After disambiguation, a second set of rules annotates discourse units as being part of the reason or the result portion of a causal relation. One aspect of deliberation is the assumption that participants in a negotiation justify their positions. Therefore, in this paper, we analyze causal relations within a single utterance of a speaker, i.e., causal relations that are expressed in a sequence of clauses which a speaker utters without interference from another speaker. As a consequence, the annotation system does not take into account causal relations that are split up between utterances of one speaker or utterances of different speakers.

Nevertheless, the reason and result portion of a causal relation can extend over multiple EDUs/sentences and this means that not only EDUs which contain the connector itself are annotated, but preceding/following units that are part of the causal relation also have to be marked. This involves deep linguistic knowledge about the cues that delimit or license relations, information which is encoded in a set of heuristics that feed the 20 different annotation rules and mark the relevant units. An example for a (simplified) relation annotation is given in (7).

(7) IF *result connector* not in first EDU of sentence AND
result connector not preceded by other connector within same sentence THEN
mark every EDU from sentence beginning to current EDU with **reason**.
ELSIF *result connector* in first EDU of sentence THEN
mark every EDU in previous sentence with **reason** UNLESS
encountering another connector.

5 Evaluation

The evaluation is split into two parts. On the one hand, we evaluate the inter-annotator agreement between five, minimally trained annotators (§5.2). On the other hand, we evaluate the rule-based annotation system against this hand-crafted gold-standard (§5.3). Each evaluation is again split into two parts: One concerns the successful identification of the causal connectors. The other concerns the identification of the spans of multilog that indicate a result/conclusion vs. a reason.

5.1 Data

The underlying data comprises of two data sets, the development and the test set. The development set, on which the above-mentioned heuristics for disambiguation and relation identification are based, consists of the transcribed protocols of the Stuttgart 21 arbitration process (henceforth: S21). This public arbitration process took place in 2010

[3] Two reviewers expressed interest in being able to access our full set of rules. Their reasons were two-fold. For one, sharing our rules would benefit a larger community. For another, the reviewers cited concerns with respect to replicability. With respect to the first concern, we will naturally be happy to share our rule set with interested researchers. With respect to the second concern, it is not clear to us that we have understood it. As far as we can tell, what seems to be at the root of the comments is a very narrow notion of replicability, one which involves a freely available corpus in combination with a freely available automatic processing tool (e.g., a machine learning algorithm) that can then be used together without the need of specialist language knowledge. We freely admit that our approach requires specialist linguistic training, but would like to note that linguistic analysis is routinely subject to replicability in the sense that given a set of data, the linguistic analysis arrived at should be consistent across different sets of linguists. In this sense, our work is immediately replicable. Moreover, given the publically available S21 data set and the easily accessible and comprehensive descriptions of German grammar, replication of our work is eminently possible.

and was concerned with a railway and urban development project in the German city of Stuttgart. The project remains highly controversial and has gained international attention. In total, the transcripts contain around 265.000 tokens in 1330 utterances of more than 70 participants.[4]

The test set is based on different, but also transcribed natural speech data, namely on experiments simulating deliberative processes for establishing a governmental form for a hypothetical new African country.[5] For testing, we randomly collected utterances from two versions of the experiment. Each utterance contained at least two causal discourse connectors. In total, we extracted 60 utterances with an average length of 71 words. There are a total of 666 EDUs and 105 instances of the markers in Table 1. The composition of the test set for each (possible) connector is in Table 2.

Reason 'because of'		**Result** 'due to'	
da	23	daher	10
weil	17	darum	11
denn	17	deshalb	12
zumal	4	deswegen	11
Total:	61		44

Table 2: Structure of the evaluation set

For the creation of a gold standard, the test set was manually annotated by two linguistic experts. 238 out of 666 EDUs were marked as being part of the reason of a causal relation, with the result/conclusion contributed by 180 EDUs. Out of 105 connectors found in the test set, 87 have a causal usage. In 18 cases, the markers have other functions.

5.2 Inter-annotator agreement

The task for the annotators comprised of two parts: First, five students (undergraduates in linguistics) had to decide wether an occurence of one of the elements in Table 1 was a causal marker or not. In a second step, they had to mark the boundaries for the reason and result/conclusion parts of the causal relation, based on the boundaries of the automatically generated EDUs. Their annotation choice was not restricted by, e.g., instructing them to choose a 'wider' or more 'narrow' span when in doubt. These tasks served two purposes: On the one hand, we were able to evaluate how easily causal markers can be disambiguated from their other usages and how clearly they introduce either the reason or the result/conclusion of a causal relation. On the other hand, we gained insights into what span of discourse native speakers take to constitute a result/conclusion and cause/reason.

For calculating the inter-annotator agreement (IAA), we used Fleiss' kappa (Fleiss, 1971), which measures the reliability of the agreement between more than two annotators. In the disambiguation task, the annotators' kappa is $\kappa = 0.96$ ("almost perfect agreement"), which shows that the annotators exhibit a high degree of confidence when differentiating between causal and other usages of the markers. When marking whether a connector annotates the reason or the result/conclusion portion of a causal relation, the annotators have a kappa of $\kappa = 0.86$. This shows that not only are annotators capable of reliably disambiguating connectors, they are also reliably labeling each connector with the correct causal relation.

In evaluating the IAA of the spans, we measured three types of relations (reason, result and no causal relation) over the whole utterance, i.e. each EDU which is neither part of the result nor the reason relation was tagged as having no causal relation. We calculated four different κ values: one for each relation type (vs. all other relation types), and one across all relation types. The IAA figures are summarized in Table 3: For the causal relation types, κ_{Reason}=0.86 and κ_{Result}=0.90 indicate near-perfect agreement. κ is significantly higher for causal EDUs than for non-causal (i.e., unmarked) EDUs ($\kappa_{\text{Non-causal}}$=0.82); this is in fact expected since causal EDUs are the marked case and are thus easier to identify for annotators in a coherent manner.

	IAA
κ_{Reason}	0.86
κ_{Result}	0.90
$\kappa_{\text{Non-causal}}$	0.82
κ_{All}	0.73

Table 3: IAA of span annotations

Across all relation types, κ_{All}=0.73 indicates "substantial agreement". The drop in the agreement is anticipated and mirrors the problem that

[4]The transcripts are publicly available for download under `http://stuttgart21.wikiwam.de/Schlichtungsprotokolle`

[5]These have been produced by our collaborators in political science, Katharina Holzinger and Valentin Gold.

is generally found in the literature when evaluating spans of discourse relations (Sporleder and Lascarides, 2008). First, measuring κ_{All} involves three categories, whereas the other measures involve two. Second, a preliminary error analysis shows that there is substantial disagreement regarding the extent of both reason and result spans. The examples in (8)–(9) illustrate this. While annotator 1 marks the result span (indicated by the *(_S* tag) as starting at the beginning of the sentence, annotator 2 excludes the first EDU from the result span.[6] In such cases, we thus register a mismatch in the annotation of the first EDU.

Nevertheless, the numbers indicate a substantial agreement. We thus conclude that the task we set the annotators could be accomplished reliably.

5.3 System performance

In order to evaluate the automatic annotation system described in Section 4, we match the system output against the manually-annotated gold standard, calculating precision, recall and (balanced) f-score of the annotation. For the disambiguation of the connectors in terms of causal versus other usages, the system performs as shown in Table 4 (the ø indicates the average of both values).

	Precision	Recall	F-score
Causal	1	0.94	0.97
Non-causal	0.85	1	0.92
ø	**0.93**	**0.97**	**0.95**

Table 4: Causal marker disambiguation

This result is very promising and shows that even though the development data consists of data from a different source, the patterns in the development set are mirrored in the test set. This means that the genre of the spoken exchange of arguments in a multilog does not exhibit the differences usually found when looking at data from different genres, as Mulkar-Mehta et al. (2011a) report when comparing newspaper articles from finance and sport.

For evaluating the annotated spans of reason and result, we base the calculation on whether an EDU is marked with a particular relation or not, i.e. if the system marks an EDU as belonging to the reason or result part of a particular causal marker and the gold standard encodes the same information, then the two discourse units match. As a consequence, spans which do not match perfectly, for example in cases where their boundaries do not match, are not treated as non-matching instances as a whole, but are considered to be made up of smaller units which match individually. Table 5 shows the results.

	Precision	Recall	F-score
Reason	0.88	0.75	0.81
Result	0.81	0.94	0.87
ø	**0.84**	**0.84**	**0.84**

Table 5: Results for relation identification

These results are promising insofar as the detection of spans of causal relations is known to be a problem. Again, this shows that development and test set seem to exhibit similar patterns, despite their different origins (actual political argumentation vs. an experimental set-up). In the following, we present a detailed error analysis and show that we find recurrent patterns of mismatch, most of which can in principle be dealt with quite straightforwardly.

6 Error analysis

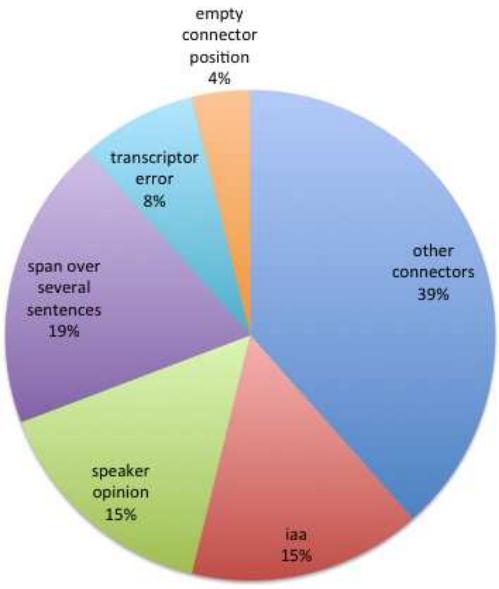

Figure 1: Error analysis, in percent.

Figure 1 shows a pie chart in which each problem is identified and shown with its share in the overall error occurrence. In total, the system makes 26 annotation errors. Starting from the top, *empty connector position* refers to structures which an annotator can easily define as reason/result, but which do not contain an overt connector. This causes the automatic annotation sys-

[6] We use the | sign to indicate EDU boundaries.

(8) Annotator 1:
(_S Ich möchte an dieser Stelle einwerfen, | dass die Frage, ob ...
I would like.Pres.1.Sg at this point add.Inf that the question if ...
'I'd like to add at this point that the question if...

(9) Annotator 2:
Ich möchte an dieser Stelle einwerfen, | (_S dass die Frage, ob ...
I would like.Pres.1.Sg at this point add.Inf that the question if ...
'I'd like to add at this point that the question if...

tem to fail. The group of *other connectors* refers to cases where a non-causal connector (e.g., the adversative conjunction *aber* 'but') signals the end of the result/conclusion or cause span for a human annotator. The presence of these other connectors and their effect is not yet taken into account by the automatic annotation system. The error group *iaa* refers to the cases where we find a debatable difference of opinion with respect to the length of a span. *Speaker opinion* refers to those cases where a statement starts with expressions like "I believe / I think / in my opinion etc.". These are mostly excluded from a relation span by human annotators, but (again: as of yet) not by the system. *Span over several sentences* refers to those cases where the span includes several sentences. And last, but not least, since the corpus consists of spoken data, an external *transcriptor* had to transcribe the speech signal into written text. Some low-level errors in this category are missing sentence punctuation. The human annotators were able to compensate for this, but not the automatic system.

Roughly, three groups of errors can be distinguished. Some of the errors are relatively easy to solve, by, e.g., adding another class of connectors, by adding expressions or by correcting the transcriptors script. A second group (*span over several sentences* and *empty connector position*) needs a much more sophisticated system, including deep linguistic knowledge on semantics, pragmatics and notoriously difficult aspects of discourse analysis like anaphora resolution.

7 Conclusion

In conclusion, we have presented an automatic annotation system which can reliably and precisely detect German causal relations with respect to eight causal connectors in multilogs in which arguments are exchanged and each party is trying to convince the other of the rightness of their stance. Our system is rule-based and takes into account linguistic knowledge at a similar level as that used by human annotators. Our work will directly benefit research in political science as it can flow into providing one measure for the deliberative quality of a multilog, namely, do interlocutors support their arguments with reasons or not?

References

James Bohman. 1996. *Public Deliberation: Pluralism, Complexity and Democracy*. The MIT Press, Cambridge, MA.

Stefanie Dipper and Manfred Stede. 2006. Disambiguating potential connectives. In *Proceedings of KONVENS (Conference on Natural Language Processing) 2006*.

Quang Xuan Do, Yee Seng Chan, and Dan Roth. 2011. Minimally Supervised Event Causality Identification. In *Proceedings of EMNLP'11*, pages 294–303.

John S. Dryzek. 1990. *Discursive Democracy: Politics, Policy, and Political Science*. Cambridge University Press, Cambridge, MA.

John S. Dryzek. 2000. *Deliberative Democracy and Beyond: Liberals, Critics, Contestations*. Oxford University Press, Oxford.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Roxana Girju. 2003. Automatic Detection of Causal Relations for Question-Answering. In *Proceedings of the ACL Workshop on Multilingual summarization and question-answering*, pages 76–83.

Amy Gutmann and Dennis Frank Thompson. 1996. *Democracy and Disagreement. Why moral conflict cannot be avoided in politics, and what should be done about it*. Harvard University Press, Cambridge, MA.

Jürgen Habermas. 1981. *Theorie des kommunikativen Handelns*. Suhrkamp, Frankfurt am Main.

Katharina Holzinger and Claudia Landwehr. 2010. Institutional determinants of deliberative interaction. *European Political Science Review*, 2:373–400.

Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press, Cambridge, Mass.

Rutu Mulkar-Mehta, Andrew S. Gordon, Jerry Hobbs, and Eduard Hovy. 2011a. Causal markers across domains and genres of discourse. In *The 6th International Conference on Knowledge Capture*.

Rutu Mulkar-Mehta, Christopher Welty, Jerry R. Hoobs, and Eduard Hovy. 2011b. Using granularity concepts for discovering causal relations. In *Proceedings of the FLAIRS conference*.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of ACL-IJCNLP*, pages 13–16.

Livia Polanyi, Chris Culy, Martin van den Berg, Gian Lorenzo Thione, and David Ahn. 2004. Sentential structure and discourse parsing. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, pages 80–87.

Rashmi Prasad and Aravind Joshi. 2008. A Discourse-based Approach to Generating Why-Questions from Texts. In *Proceedings of the Workshop on the Question Generation Shared Task and Evaluation Challenge*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of LREC 2008*, pages 2961–2968.

Ted Sanders. 2005. Coherence, Causality and Cognitive Complexity in Discourse. In *Proceedings of SEM-05, First International Symposium on the Exploratiaon and Modelling of Meaning*, pages 105–114.

Angela Schneider and Manfred Stede. 2012. Ambiguity in German Connectives: A Corpus Study. In *Proceedings of KONVENS (Conference on Natural Language Processing) 2012*.

Caroline Sporleder and Alex Lascarides. 2008. Using Automatically Labelled Examples to Classify Rhetorical Relations: An Assessment. *Natural Language Engineering*, 14(3):369–416.

Manfred Stede. 2004. The Potsdam Commentary Corpus. In *In Proceedings of the ACL'04 Workshop on Discourse Annotation*, pages 96–102.

Yannick Versley and Anna Gastel. 2012. Linguistic Tests for Discourse Relations in the Tüba-D/Z Corpus of Written German. *Dialogue and Discourse*, 1(2):1–24.

Yannick Versley. 2010. Discovery of Ambiguous and Unambiguous Discourse Connectives via Annotation Projection. In *Workshop on the Annotation and Exploitation of Parallel Corpora (AEPC)*.