

Annotating Discourse Relations in Spoken Language: A Comparison of the PDTB and CCR Frameworks

Ines Rehbein, Merel Scholman, Vera Demberg

Saarland University
Campus C7.4, 66123, Saarbrücken
{rehbein, m.c.j.scholman, vera}@coli.uni-saarland.de

Abstract

In discourse relation annotation, there is currently a variety of different frameworks being used, and most of them have been developed and employed mostly on written data. This raises a number of questions regarding interoperability of discourse relation annotation schemes, as well as regarding differences in discourse annotation for written vs. spoken domains. In this paper, we describe our work on annotating two spoken domains from the SPICE Ireland corpus (telephone conversations and broadcast interviews) according to two different discourse annotation schemes, PDTB 3.0 and CCR. We show that annotations in the two schemes can largely be mapped onto one another, and discuss differences in operationalisations of discourse relation schemes which present a challenge to automatic mapping. We also observe systematic differences in the prevalence of implicit discourse relations in spoken data compared to written texts, and find that there are also differences in the types of causal relations between the domains. Finally, we find that PDTB 3.0 addresses many shortcomings of PDTB 2.0 wrt. the annotation of spoken discourse, and suggest further extensions. The new corpus has roughly the size of the CoNLL 2015 Shared Task test set, and we hence hope that it will be a valuable resource for the evaluation of automatic discourse relation labellers.

Keywords: Annotation of discourse relations (DRs), interoperability of annotation schemes, DRs in spoken and written genres

1. Introduction

Over the last decade, research in NLP has widened its scope, moving beyond the sentence level to analysing the discourse structure of a text. This has resulted in the creation of discourse-annotated corpora, such as the Penn Discourse Treebank (PDTB) (Prasad et al., 2008), the Rhetorical Structure Theory (RST) Treebank (Carlson et al., 2002), and the Annodis corpus (Segmented Discourse Representation Theory, SDRT) (Afantenos et al., 2012), to name a few. However, as of yet, the different frameworks are not inter-operable, nor is there a unified scheme for discourse annotation (but see the proposals by Benamara and Taboada (2015), Chiarcos (2014), Bunt et al. (2012), and Sanders et al. (In preparation)).

Most discourse-annotated corpora are based on written rather than spoken text. This point is crucial, as spoken and written texts are produced and processed differently (Cuenca, 2015): Spoken communication is characterised by a high degree of interactivity that requires turn-taking devices for discourse management; sentence length on average is shorter, and the pressure of rapid online processing often leads to disfluent structures. In contrast to written communication, the speaker and the hearer have access to additional channels of communication, such as visual information or, at least, audio cues such as pitch and sentence stress, and we observe many elliptical structures and omissions. We expect that these differences will be reflected in the use of discourse relations (DRs) in the spoken domain. The above considerations raise the question whether annotation schemes developed for written language are adequate for describing coherence relations in spoken language. Even for written text, there is no consensus on which and how many categories of coherence relations should be distinguished. Most proposals agree that coherence relations are binary relations between two discourse

elements, but they differ in their operationalisation of how to annotate these relations. RST, for example, assumes that the appropriate representation for discourse relations is a tree, while other frameworks do not make the same assumption.

The present paper focusses on two annotation frameworks, namely PDTB 3.0 (Prasad et al., In preparation) and the Cognitive approach to Coherence Relations (CCR; Sanders et al. (1992)). These schemes differ greatly in how they describe and annotate coherence relations: PDTB distinguishes end labels (Figure 1), i.e. Contingency.Cause.Result, whereas CCR describes relations according to four cognitive dimensions (*polarity, basic operation, source of coherence, order*), i.e. a positive causal objective forward relation (Example 1).¹

(1) *Her flight was late, so she missed her connection.*

a. PDTB: CONTINGENCY.CAUSE.RESULT

b. CCR: *positive causal objective forward*

We present an annotation experiment where the two frameworks are applied to English data from the spoken domain. We are interested in the following questions:

1. In what sense are annotations dependent on the formalism chosen? Do we obtain equivalent information? What possible biases are introduced by the annotation guidelines? (Section 5.)
2. What are the differences between discourse relations in the spoken and in the written domain? (Section 6.)
3. Can the two frameworks adequately describe coherence relations in spoken language, or do we need additional categories? (Section 7.)

¹In all our examples, we encode the first argument in italics and the second argument in bold face. Explicit discourse connectives are underlined.

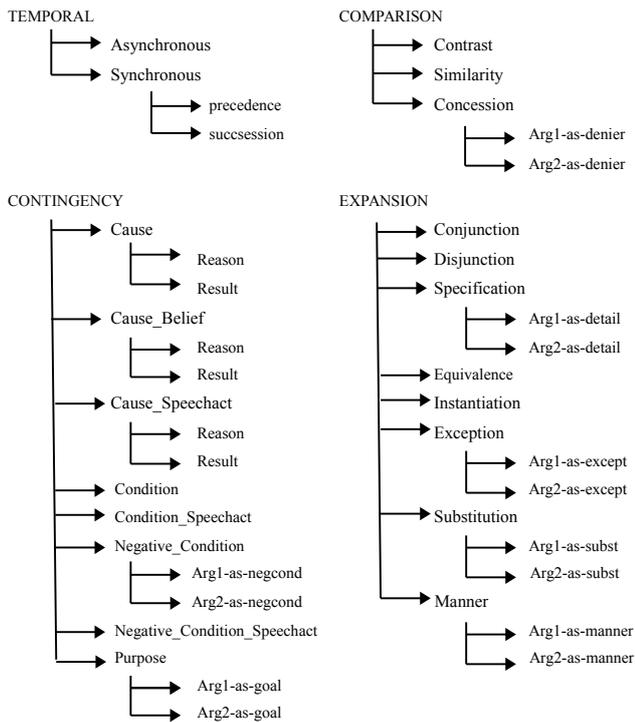


Figure 1: PDTB 3.0 sense hierarchy

2. Discourse annotation frameworks

2.1. PDTB

The Penn Discourse Treebank 2.0 (Prasad et al., 2008) is a resource of discourse relation annotations on texts from the Wall Street Journal corpus. The PDTB framework aims at describing local links between arguments, rather than building up a full tree as a representation of the entire text. Each relation is therefore analysed independently of other relations. The PDTB taxonomy was originally developed for written language, but PDTB 2.0 has also been applied to spoken data. This will be discussed further in Section 3.

The segments that make up a discourse relation are labeled Arg1 and Arg2 in PDTB. These labels are assigned based on syntactic restrictions: Arg2 is the argument that appears in the clause that is syntactically bound to the connective. Each relation is assigned an end label based on a three-tiered hierarchy of relation senses. The top-level, or *class*, consists of four major semantic classes, namely TEMPORAL, CONTINGENCY, COMPARISON, and EXPANSION. Each level has been divided into multiple sublevels named *types*, which in turn can be divided into *subtypes*. Annotators can choose to annotate only at coarse-grained senses if they have low confidence on the fine-grained senses such as the subtype.

In the current study, spoken discourse has been analysed according to the PDTB 3.0 sense hierarchy (Figure 1), a revised version of PDTB 2.0. For an overview of changes in the framework, see Prasad et al. (In preparation).

2.2. CCR

The annotation framework of CCR is based on a cognitive theory of coherence relations, proposed by Sanders et al. (1992). The theory has since been used to annotate various types of discourse, including spoken text, chat frag-

ments, and children’s language use (see Sanders, Vis & Broeder, (2012) and Sanders et al., (In preparation), for an overview). CCR differs from other annotation frameworks in that coherence relations are not assigned a specific end label, but rather are defined by their characteristics. In CCR, four cognitive dimensions are distinguished that apply to every relation. These dimensions are polarity, basic operation, source of coherence, and order of the segments. They will be explained shortly in the next paragraphs, but for a more detailed explanation, see Sanders et al. (1992) and Scholman et al. (2016).

The polarity of a relation refers to the positive or negative character of a segment. A relation is positive if the propositions P and Q, expressed in the two discourse segments S1 and S2, are linked directly, without a negation of one of these propositions. A relation with a positive polarity is typically connected by connectives such as *and* or *because*. A relation is negative if the negative counterpart of either P or Q functions in the relation. A relation with a negative polarity is typically connected by connectives such as *but* and *although*.

The basic operation distinguishes between causal and additive relations. A relation is causal if an implication relation ($P \rightarrow Q$) can be deduced between the two segments. Causal relations are typically connected by *because* and *so*. The category of causal relations also comprises conditional relations (cf. Scholman et al., 2016). A relation is additive if the segments are connected as a conjunction ($P \& Q$). Temporal relations are considered a subclass of additive relations, in which their segments are ordered in time.

The third dimension, the Source of Coherence, distinguishes between objective and subjective relations. A relation is subjective if the author or speaker is actively engaged in the construction of the relation. Subjective relations express the speaker’s opinion, argument, claim, or conclusion. Objective relations, on the other hand, consist of segments that describe situations that occur in the real world.

The fourth dimension is the order of the segments. This dimension applies to causal, conditional and temporal relations. In a coherence relation with a basic order, the antecedent (P) is S1, followed by the consequent (Q) as S2. In a relation with a non-basic order, P maps onto S2 and Q onto S1. For causal and conditional relations, P is the cause, condition or argument, and Q is the consequence or the claim. For temporal relations, P is the first event, and Q is the second event.

3. Related Work

Two different strands of research are relevant to our work. The first one focusses on the unification of different annotation schemes for the annotation of discourse relations, the second is on developing or adapting discourse annotation schemes for spoken language.

3.1. Mapping annotations across frameworks

Bunt et al. (2012) and Prasad and Bunt (2015) describe efforts to create an international standard for the annotation of discourse with semantic relations, based on different frameworks for discourse annotation. They define a set of 20 core

discourse relations (Prasad and Bunt, 2015) that they consider to be indispensable for the annotation of DRs, together with clear definitions and examples for each relation. In future work, they plan to provide mappings from their core set to the annotation labels in different DR frameworks such as the PDTB, RST, SDRT and CCR.

Such an ambitious enterprise raises questions about the inter-operability of existing annotation frameworks for discourse. Chiarcos (2014) explores the inter-operability of the RST and PDTB frameworks, with the goal of extending the Ontologies of Linguistic Annotation (OLiA) (Chiarcos and Sukhareva, 2012) with discourse features. With that aim in view, the labels in the different frameworks are mapped onto each other. However, the mapping does not include the structures (i.e. the arguments) of the annotated relations, and no claim is made about the completeness of this endeavour. Chiarcos observes that in order to fully map the annotations from the RST and PDTB corpora, structural transformations are necessary, e.g. a conversion of the annotations to dependency DAGs.

Benamara and Taboada (2015) propose a taxonomy for mapping discourse relations across two different frameworks, *Rhetorical Structure Theory* (RST) and *Segmented Discourse Representation Theory* (SDRT), and use it to map the annotations in three resources (the RST-DT English corpus (Carlson et al., 2002), the SDRT Annodis French corpus (Afantenos et al., 2012), and the RST Spanish Treebank (da Cunha et al., 2011)). The authors claim that their taxonomy is robust across theoretical frameworks and can be applied to corpora from different languages. However, they identify some problems for the mapping, e.g. SDRT does not distinguish between causal and epistemic uses of causal relations, and the CAUSE-RESULT and CONSEQUENCE relations in the RST-DT corpus are similar and can correspond to either REASON or RESULT.

Sanders et al. (In preparation) propose to use unifying dimensions to be able to ‘translate’ annotations from one framework to another. In their proposal, they extend the dimensions distinguished by CCR with additional criteria needed to capture more fine-grained relation senses. They then decompose the relation labels distinguished in the PDTB 2.0, PDTB 3.0, RST and SDRT frameworks according to these dimensions. Many of the relation labels can be mapped onto the dimensions automatically, but manual annotation will also be required for certain labels or dimensions.

The approaches above are similar in their aim to define a mapping between existing annotation frameworks. What is missing so far is a validation of the proposed mappings in terms of an annotation experiment. The proposed mapping can be used to predict annotations carried out on the same data. Let’s assume that we have a mapping M between relation A in framework $F1$ and relation B in framework $F2$. Every time one annotator assigns label A from framework $F1$ to a data point in the corpus, we predict that the second annotator assigns label B from framework $F2$. The mapping, that was created based on theoretical assumptions, can be considered as verified if a substantial part of the annotations in the two frameworks can be translated into each other, using the mapping.

In Section 4., we present such a validation experiment, using the CCR and PDTB 3.0 frameworks, and analyse the results with respect to the predictions made by our mapping.

3.2. Annotating DRs in spoken language

While most annotation projects have focussed on annotating discourse relations in written text, there are some studies on annotating discourse-relational devices in the spoken domain.

Tonelli et al. (2010) adapt the PDTB 2.0 annotation scheme to make it more suitable for annotating spoken data from the LUNA Corpus, which is a language resource with help-desk dialogues on the topic of software/hardware troubleshooting. The special properties of spoken conversation caused them to change the PDTB sense hierarchy and include new relations. The most important change is the addition of speechact relations in the sense of Sweetser (1990). Other changes include the GOAL relation, likely motivated by the peculiarities of the task-oriented help-desk dialogues. They also discarded the LIST relation, as it never occurred in their corpus of conversational speech.

These changes to the PDTB hierarchy are also reflected in the revised hierarchy of the PDTB 3.0 (Prasad et al., In preparation) (i.e. the addition of the level-2 sense PURPOSE with level-3 subsenses ARG1-AS-GOAL, ARG2-AS-GOAL, as well as the distinction between BELIEF and SPEECHACT readings).

Other studies that have used the PDTB framework to annotate spoken data are Demirşahin and Zeyrek (2014) and Stoyanchev and Bangalore (2015). However, they have not made any further changes to the PDTB framework.

4. Data & Method

The data we use in our annotation experiment comes from the SPICE-Ireland corpus (Kallen and Kirk, 2012), a corpus of spoken Irish English containing a variety of different genres, from which we selected *broadcast interviews* and *telephone conversations* for discourse relation annotation. The PDTB corpus also includes texts from different genres: *essays*, *summaries*, *news* and *letters*, as described in Webber (2009). This will allow us to investigate the differences in the use of DRs in different genres from the spoken and written domain.

SPICE-Ireland corpus The corpus includes the spoken part of the ICE-Ireland corpus (Kallen and Kirk, 2008), with texts ranging over 15 different discourse settings (e.g. broadcast discussions, broadcast news, classroom discussions, private telephone conversations, or parliamentary debates, amongst others). SPICE-Ireland comprises pragmatic annotations on top of the transcriptions from the ICE-Ireland corpus, including mark-up for speech-act functions and discourse markers. The annotation of discourse markers is restricted to modulating devices such as *well*, *you-know*, *sure*, *kind-of* and does not consider discourse connectives and coherence relations.

Annotation procedure Two linguistically trained coders annotated discourse relations in the spoken data according to the PDTB 3.0 and CCR frameworks. The discourse relations in each text were annotated by one coder following

genre	written (PDTB)				spoken (SPICE)	
	essays	summ.	letters	news	broadc.	teleph.
no. sent	6,517	1,667	911	38,963	1,507	2,717
no. words	139,445	31,316	18,207	821,104	20,801	20,239
no. DRs	6,468	933	750	32,018	1,244	1,201
explicit	45.8	17.5	37.5	37.8	44.0	25.5
implicit	42.9	15.1	33.5	32.3	27.2	12.6

Table 1: Distribution of explicit and implicit coherence relations across different genres, normalised frequencies per 100 sentences/speech units. The total number of DRs includes Explicit, Implicit, AltLex and EntRel, **implicit** only refers to the number of implicit relations, without EntRel and AltLex.

the PDTB guidelines, and by the second coder according to CCR. A subset of the data was annotated by both annotators with both frameworks, in order to determine inter-annotator agreement. In each file, all explicit discourse connectors as well as all implicit coherence relations have been encoded. We follow Tonelli et al. (2010) in not restricting ourselves to annotating implicit relations between adjacent speech units only. However, we did not annotate implicit relations between arguments uttered by different speakers.

The annotated subcorpus includes the private telephone conversation and the public broadcast interview subsections of the SPICE-Ireland corpus. Table 1 shows the size of the subcorpora from four different genres from the Penn Treebank (Webber, 2009) (left four columns) and our newly annotated data from the SPICE corpus (right two columns). Overall, 2,445 discourse relations have been annotated according to the PDTB 3.0 framework, and 2,069 discourse relations have been encoded with CCR dimensions. The lower number of CCR dimensions is due to theoretical differences between the two frameworks. Consider, for example, the case of multiple connectives such as *but then*. According to the PDTB scheme, each connective receives a label, while in the CCR framework those cases obtain only one label. Another example are the PDTB labels ENTREL and NOREL, which are not annotated in CCR.

Inter-annotator agreement (IAA) Table 2 shows the inter-annotator agreement for discourse relation annotations in CRR and PDTB for a subset of the data double-annotated with both frameworks. In the evaluation, we only consider the senses/dimensions for instances that have received an annotation in both frameworks. For CCR, this amounts to 289 annotated discourse relations (explicit and implicit) in 4 different texts of broadcast interviews, and for the PDTB framework, we compared 175 explicit annotations on the same texts.

dimension	CCR		PDTB 3.0	
	% agr.	κ	% agr.	κ
polarity	92.0	.802	-	-
basic operation	81.3	.717	-	-
source of coherence	81.3	.631	-	-
order	86.2	.867	-	-
all	61.9	0.555	84.6	.797

Table 2: Inter-annotator agreement for coherence relations in CCR and PDTB

For the CCR framework, we observe a percentage agreement ranging between 81.3% and 92% for the individual dimensions. The source of coherence dimension shows the

lowest IAA; a result that has been found in other studies as well (e.g., Scholman et al. (2016) and Sanders et al. (1992)). The IAA for the combined dimensions (exact match: an instance counts as correct only if the annotators assigned the same values for all four dimensions) is around 62%, with a κ of 0.555.

For the PDTB framework, we achieved a percentage agreement of 84.6% for the most fine-grained sense distinctions for explicit relations. A comparison of our results to the ones for the English PDTB corpus is not straightforward as we annotated according to the revised PDTB 3.0 sense hierarchy while the agreements reported in the literature for the English Penn Discourse treebank refer to PDTB 2.0-style annotations. However, our IAA seems to be roughly in the same range as the one of Prasad et al. (2008), who report a percentage agreement of 84% for level-2 senses and of 80% for level-3 senses (for all discourse relations). Tonelli et al. (2010) do not report IAA for their corpus of spoken Italian.

5. Comparison of PDTB and CCR

The relevant questions for this section are whether the labels of the frameworks can be mapped onto each other without loss of information, and how the two frameworks influence the annotation process and the result.

Sanders et al. (In preparation) created a mapping between PDTB 3.0 and CCR, for which they decomposed every PDTB relation into a specific combination of values for CCR dimensions. This decomposition allows us to test whether coherence relations annotated according to the PDTB framework by one annotator fall into corresponding categories when analysed by another annotator using CCR. In order to test this, all PDTB annotations and CCR annotations of the SPICE-Ireland corpus have been mapped onto each other. This has resulted in the correspondence matrix shown in Table 3.² The cells with the bold-font, underlined numbers indicate the predicted mapping as proposed in the decomposition created by Sanders et al. (In preparation).

Looking at the PDTB relations in Table 3 first, it can be noted that many labels in the EXPANSION class fall into the same CCR categories: positive, additive, objective/subjective. Further distinctions will be necessary to create more fine-grained differences between these relations, for example to account for a specification relation relative to other additive relations. Sanders et al. (In preparation) have proposed several additional features that should capture these more fine-grained distinctions, such as a feature for the specificity of the relation. An additional distinction that is made in PDTB 3.0 but not in CCR is the speechact reading. In PDTB 3.0, there are separate labels for this class of relations (CONTINGENCY.CAUSE_SPEECHACT and CONTINGENCY.CONDITION_SPEECHACT), whereas in CCR, speechact relations are classified as subjective relations. A further distinction in the source of coherence would thus account for speechact relations.

²For a complete version of the table, including level-three distinctions, see http://www.sfb1102.uni-saarland.de/?page_id=2582.

<i>Polarity</i>		pos	neg	neg	neg	neg	pos	pos								
<i>Basic op.</i>		temp	temp	temp	caus	caus	caus	caus	cond	caus	caus	add	add	add	add	count
<i>S. of coh.</i>		obj	obj	obj	obj	obj	subj	subj	obj	subj	obj	subj	obj	subj	obj	subj
<i>Order</i>		na	forw	back	forw	back	forw	back	undsp	undsp	undsp	undsp	na	na	na	na
Temp.	Synchronous	68	13	4	0	0	0	0	0	0	0	0	2	8	5	53
	Asynchronous	3	67	9	3	0	1	2	1	0	0	3	2	6	3	105
Cont.	Cause	0	2	0	17	12	25	36	1	0	0	0	1	3	3	300
	Cause_belief	0	0	0	5	5	40	32	0	0	0	0	0	9	9	22
	Cause_speechact	0	0	0	0	0	53	47	0	0	0	0	0	0	0	15
	Condition	3	1	1	0	0	0	1	30	58	0	0	0	0	1	77
	Condition_speechact	0	0	0	0	0	0	0	0	93	0	0	0	0	0	14
Comp.	Concession	0	2	0	0	0	4	0	0	2	10	25	20	32	2	56
	Contrast	0	1	0	0	1	0	1	0	0	2	10	33	43	2	206
	Disjunction	0	0	0	0	0	0	0	0	0	0	0	25	55	0	20
Exp.	Substitution	0	0	0	0	0	0	0	0	0	7	0	22	64	7	14
	Conjunction	1	16	1	2	0	6	2	0	0	0	0	4	6	31	538
	Equivalence	0	0	0	0	2	4	32	0	0	0	0	2	14	42	45
	Instantiation	0	0	2	0	0	0	24	0	2	0	0	3	3	16	38
	Specification	0	0	1	1	6	6	23	0	1	0	1	3	4	24	143

Table 3: Distribution (%) of explicit and implicit relations, only labels and categories where $n > 10$ (undsp: underspecified, na: not applicable, nra: no relation annotated) and raw counts

Even though the PDTB 3.0 sets apart speechact relations from other relations more clearly than CCR, it does not seem to set apart other subjective relations from objective relations in a similar manner as CCR. Looking at the categories that CCR distinguishes but the PDTB doesn't, it is revealed that the PDTB often does not distinguish between the source of coherence in its relation labels. As a result, most relation labels are still underspecified regarding their source of coherence. The same goes for CCR's order: PDTB has encoded Arg1-Arg2 order in their labels, but does not take into account the surface order of their arguments. As a result, the order of the segments remains underspecified for most PDTB labels.

Ideally, each type of relation distinguished in one framework would be annotated as a similar type in the other framework; i.e. all relations with the PDTB label CONTINGENCY.CAUSE should be annotated as positive causal relations according to the CCR dimensions. A look at Table 3 reveals that this is not always the case for every relation.

Overall, 70% of the PDTB relations were consistently categorised as belonging to the target CCR class. For relations where the PDTB label does not map to the predicted CCR categories, the mismatch could be due to different factors. It could be due to different interpretations by the annotators (as reflected also in interannotator agreement when annotating within a single framework), or it might be caused by differences in the annotation process and discourse relation definitions of the two frameworks. For example, CCR advises annotators to use connective substitution and paraphrase tests (cf. Knott and Sanders (1998), Scholman et al. (2016)), whereas PDTB does not make use of such tests to determine the label for an explicit relation.

The issues described above, and especially the fact that the frameworks differ in granularity of distinctions between discourse relations, constitute challenges for the goal of achieving a one-to-one mapping. Our suggested mapping is therefore a many-to-many mapping between DRs from the PDTB and CCR frameworks. Finally, there can be differences between the frameworks in what is annotated as a relation. For example, PDTB annotates ENTREL relations, whereas CCR does not annotate such cases. In the current mapping, we therefore only consider relations other than ENTREL and NOREL.

To determine whether the cases in which the label in one

framework did not map onto the predicted label in another framework are due to differences between the frameworks (e.g. one theory having a more restricted definition of a specific label) or are caused by other factors, a random sample of 50 disagreements was analysed. 19 disagreements of the 50 were due to annotation errors, i.e. after discussing the relation, one annotator agreed that another label was more appropriate, and 7 disagreements were caused by differences in segmentation or interpretation. For example, the annotators had annotated a different relation and could agree to both values according to a different interpretation. The remaining 24 cases of the sample of 50 disagreements were in fact caused by differences in operationalisation of the frameworks. This means that the agreement in PDTB 3.0 and CCR annotations for the many-to-many mapping used here may be even higher than the 70%.

Most of the mismatches due to annotation operationalisations can be classified as one of three categories. The first category of disagreements focusses on the definition of CONCESSION in PDTB and negative causals in CCR. CONCESSIONS are mapped as negative causal relations, but in our experiment, the relations that one annotator classified as COMPARISON.CONCESSION in PDTB were often classified as negative additive by the other annotator (in 52% of the cases). In PDTB, the class CONCESSION is described as containing relations for which one argument creates an expectation that the other denies. The annotator needs to decide what constitutes an expectation. In CCR, negative causals are described as relations between P and the negative counterpart of Q. A suggested substitution test to determine whether a negative relation is additive or causal, is to reverse the polarity and then determine the basic operation. This is illustrated in Example 2, which is taken from a fragment of a speaker who wants to go to a party, but needs to work in a shop instead.

- (2) Original:
Cos I wouldn't mind going down. However the shop won't be closed I'd say until about seven.
 (And that's a bit too late really to go to Derry.)

Substitution:
Cos I wouldn't mind going down. So/Because the shop will be closed around seven.

In PDTB terms, the first argument leads to the expectation that the speaker will go to the party, which is then denied in the second argument. This would therefore be labeled as a CONCESSION.CONTRA-EXPECTATION. The substitution version is however not the same relation as the original relation: it becomes clear that the second argument of the relation is not the true cause or the result of the first argument. It would therefore not be labeled as a causal negative relation according to CCR. The substitution test could thus result in a stricter definition of what can be considered a negative causal relation, which in turn causes disagreement between the frameworks. It is unclear, however, whether this is really a difference between the frameworks, or whether it is due to annotator bias.

The second category of disagreements revolves around argumentative relations. The CCR dimension source of coherence distinguishes between objective and subjective relations, such as cause-consequence (objective causal relations) and argument-claim (subjective causal relations). Hence, in CCR, if an implicit relation in which one segment provides an argument for the other can be connected by *because*, the relation can be classified as a causal subjective relation. In PDTB, however, such relations can be classified as additive EXPANSION relations, mainly belonging to the types INSTANTIATION, SPECIFICATION and EQUIVALENCE. Example 3 illustrates this issue.

- (3) *I used the weight room facility for exercising.* (implicit because) **I exercise from physiotherapy that I had to do.**
- a. PDTB: EXPANSION.SPECIFICATION
 - b. CCR: positive causal subjective backward

The third category of disagreements centers around additive negative relations. In order to determine whether a coherence relation is negative, CCR makes use of a substitution test: if the two arguments of a relation can be connected by *but*, this indicates that the relation is negative. PDTB does not follow a similar guideline, thereby not automatically classifying these relations as belonging to the COMPARISON class. Example 4 illustrates this issue.

- (4) *She's by a Northern-based sire,* (implicit but) **I think he's dead now perhaps.**
- a. PDTB: Expansion.Specification
 - b. CCR: negative additive subjective (order not applicable)

The main cause of disagreements for these last two categories thus seems to be the relative importance of connectives: CCR relies on specific connectives in a specific order of importance (with *but* being the strongest and *and* being the weakest indicator) to determine the values of the dimensions, whereas PDTB does not rely on connectives as heavily to determine the end labels.

In sum, the mapping of PDTB and CCR annotations has shown that PDTB relations fall into the target CCR class relatively often (70% of the cases). When a relation does not fall into the target CCR class, the disagreement is often due to (a) annotator errors, irrespective of the frameworks, (b) differences in the definition of what constitutes

a negative causal relation, (c) the operationalisation of the annotation process leading to different annotations, or (d) differences in the segmentation rules leading to the annotation of relations in one framework that the other does not allow (ENTRREL AND NOREL).

6. Genre differences

As pointed out by Webber (2009) based on a discourse analysis of different newspaper genres from the Wall Street Journal, text genre crucially affects the distribution of discourse relations. For our corpus, we observe a substantial difference in the frequency of coherence relations between the spoken and written genres, as well as between the two spoken genres (Table 1). In this section, we will therefore focus on both comparisons respectively.

6.1. Spoken vs. written genres

Comparing the spoken to the written genres, the most striking difference is the proportion of implicit to explicit discourse relations. In the written genres, we have roughly the same proportion (with slightly more explicit relations). In the spoken language data, explicit relations are about twice as frequent as the implicit ones (Table 1). This observation is consistent with Tonelli et al. (2010) who report that explicit relations were more than twice as frequent as implicit ones in the spoken LUNA corpus.

Another difference concerns the use of causal relations in the data. In written genres, situations described in the first segment are oftentimes direct causes of situations described in the second segment, see e.g. Example (5). In spoken language however we find many examples where the causal link between the two arguments is not very strong. Instead, *so* often marks a conclusion based on the information in the first segment (Example 6).

- (5) *but times have changed, even in Utah* so **Mr. Redford no longer stands out as an extremist**
- (6) *I've already had a meeting uhm an update meeting* so **the place hasn't burnt down or anything**

We annotated those instances as CAUSE.BELIEF, following the revised version of the PDTB 3.0 (PRAGMATIC CAUSE in PDTB 2.0). However, in the PDTB 2.0 corpus no instances have been found for the subtype PRAGMATIC CAUSE.JUSTIFICATION, where the second argument expresses the claim and the first argument the justification, as in example (6) above. In our corpus of spoken data, we found 13 instances of this particular type.

6.2. Broadcast interviews vs. telephone conversations

To get a better idea of the differences between the two spoken genres, namely private telephone conversations and public broadcast interviews, we look at the distribution of individual coherence relations in the data. Table 4 shows the most frequent level-2 relations in the two data sets.

One obvious difference is the higher amount of causal relations in the telephone conversations. Especially for the implicit relations the gap is quite high: a more detailed analysis of subtypes of causal relations shows that 8% (BC) versus 19% (TEL) of all implicit relations are annotated as CONTINGENCY.CAUSE.REASON, and 4.4%

	Broadcast		Telephone	
	exp	imp	exp	imp
Temporal.Asynchronous	7.4	1.5	11.3	3.2
Temporal.Synchronous	5.3	0.2	3.9	0.0
Contingency.Cause	13.0	12.5	21.1	33.8
Contingency.Cause_Belief	2.0	1.0	1.3	0.0
Contingency.Cause_Sp.act	0.0	0.7	1.1	3.2
Contingency.Cond	7.8	0.0	7.1	0.0
Contingency.Cond_Sp.act	3.0	0.0	0.5	0.0
Comparison.Concession	4.5	2.0	4.5	1.6
Comparison.Contrast	13.7	8.3	12.3	11.4
Expansion.Conjunction	32.3	29.8	32.9	13.9
Expansion.Equivalence	0.9	11.5	0.2	10.4
Expansion.Instantiation	1.7	7.3	0.3	0.9
Expansion.Specification	4.7	23.5	0.3	19.9
Other	3.8	1.7	3.2	1.9
total (%)	100.0	100.0	100.0	100.0

Table 4: Distribution of the most frequent coherence relations across the spoken genres

(BC) versus 14.9% (TEL) are annotated as CONTINGENCY.CAUSE.RESULT. Despite this striking difference, in both genres the majority of explicit causal relations are marked by the same connectives, *because* and *so*.

7. Challenges for the annotation of DRs in spoken language

Next, we will discuss challenges for the annotation of discourse relations in spoken language. Example (6) illustrates a Conclusion relation that is rare in written language. While no annotated instances have been found in the PDTB 2.0 corpus, the PDTB 3.0 annotation scheme does provide the labels to deal with those cases (CONTINGENCY.CAUSE_BELIEF.RESULT). In CCR, on the other hand, examples like this would be annotated as *positive causal objective forward* and thus would not be distinguishable from other subjective causal relations such as Examples (1) or (5). Further distinctions would be necessary to account for such relations.

Spoken language also provides additional means to express contrast, e.g. through topicalisation and sentence stress (see Example (7)), that are not easily accessible in the written medium.³ These instances are comparable to the ones annotated as AltLex in the PDTB, where the existence of an alternative lexicalisation blocks the insertion of an implicit connective (compare 8a, 8b). One solution is to introduce new categories similar to AltLex, e.g. AltTop and AltStress (Alternative Topicalisation, Alternative Stress).

- (7) Context: Are you busy over Halloween?
Uhm I'm not actually because oh-well THAT week you mean
- (8) *I'm not busy in the first week of October*
 a. but you mean **THAT week**
 b. *but **THAT week you mean**

Another phenomenon typical for spoken discourse is the repetition of identical or near-identical sequences, often used for reinforcement or intensification (examples (9) and

(10)). We annotated those cases as EQUIVALENCE. However, this might be in contrast to the originally intended meaning of the label in the PDTB, where EQUIVALENCE was used when the same situation is described from different perspectives (Prasad et al., In preparation).

- (9) SPK1: He 's okay
 SPK1: he 's okay
- (10) SPK1: That 's brilliant
 SPK1: Oh fantastic news
 SPK1: Well done Rachel-Anne
 SPK1: That 's brilliant

Similar to Tonelli et al. (2010), we decided against limiting the annotation of implicit DRs to adjacent sentences, as done for the English PDTB corpus, as in spoken discourse arguments are often separated by fragments or disfluent segments (e.g. unfinished utterances, backchannel signals, exclamations; see example (11)).

- (11) SPK1: I 'm on email every day_{ARG1}
 SPK1: you know
 SPK1: I can
 SPK1: I 've access to it now_{ARG2}

Overall, the issues discussed here need to be addressed in the annotation guidelines of the two frameworks. The revised version of the PDTB (Prasad et al., In preparation) already addresses some properties of spoken language. For example, in our data we found frequent uses of pragmatic discourse relations, both epistemic and speechact, thus confirming the usefulness of the revised PDTB 3.0 hierarchy.

8. Conclusions

In the present paper, we described our annotation of coherence relations with respect to two frameworks, PDTB 3.0 and CCR, on two spoken domains, private telephone conversations and public broadcast interviews. The contributions of this paper are three-fold.

Firstly, the present annotation effort is the first to annotate a single larger amount of text with both CCR and PDTB, and analyse in detail the agreement of two discourse relation annotation frameworks on the same text. Our most important findings are that PDTB 3.0 and CCR annotations can be mapped onto one another with high confidence for most relations: 70% of the relations with a PDTB label fell into the predicted CCR categories. An analysis of a subset of the remaining 30% revealed that half of the label mismatches were caused by annotator disagreement which was due to differences in interpretation and independent of the frameworks. Based on this evaluation, the agreement between labels/categories of PDTB 3.0 and CCR can therefore be estimated to be even higher than 70%. Qualitatively different annotations result from differences in the annotation guidelines, in particular with respect to tests designed to help the annotator to decide for a discourse relation, e.g. by connective insertion tests. The discourse relation frameworks also differ in "granularity", i.e. one framework makes distinctions that are not reflected in categories from the other framework. This leads to a many-to-many mapping between discourse relation labels.

³In computer-mediated communication, capitalisation can be used to mark stress.

Secondly, we have identified differences in distribution of discourse relations between genres. Most strikingly, we observed a higher proportion of explicit to implicit relations in spoken data compared to written data: explicit relations are approximately twice as frequent as implicit relations in spoken data. In written genres, on the other hand, the proportion is roughly the same.

Finally, we have discussed challenges for applying the annotation schemes to spoken language. PDTB 3.0 has addressed many issues that were present in PDTB 2.0, for example by adding separate labels for speechact relations. CCR does not distinguish between speechact relations and other subjective relations. We argue that this could be a useful distinction for the analysis of spoken data. We also identified two additional categories of relations that could be added to PDTB 3.0 for spoken data, namely Alternative Topicalisation and Alternative Stress.

9. Acknowledgements

This research was funded by the German Research Foundation (DFG) as part of SFB 1102 “Information Density and Linguistic Encoding” and the Cluster of Excellence EXC 284 “Multimodal Computing and Interaction”.

10. References

- Afantenos, S., Asher, N., Benamara, F., Bras, M., Fabre, C., Ho-Dac, M., Draoulec, A. L., Muller, P., Pery-Woodley, M.-P., Prevot, L., Rebeyrolles, J., Tanguy, L., Vergez-Couret, M., and Vieu, L. (2012). An empirical resource for discovering cognitive principles of discourse organisation: the ANNODIS corpus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation, LREC'12*.
- Benamara, F. and Taboada, M. (2015). Mapping different rhetorical relation annotations: A proposal. In *Proceedings of the 4th Joint Conference on Lexical and Computational Semantics, *SEM*.
- Bunt, H., Prasad, R., and Joshi, A. (2012). First steps towards an ISO standard for annotating discourse relations. In *Proceedings of the Joint ISA-7, SRSL-3, and I2MRT LREC 2012 Workshop on Semantic Annotation and the Integration and Interoperability of Multimodal Resources and Tools*.
- Carlson, L., Marcu, D., and Okurowski, M. E. (2002). RST discourse treebank, ldc2002t07. Web Download. Philadelphia: Linguistic Data Consortium.
- Chiarcos, C. and Sukhareva, M. (2012). Ontologies of linguistic annotation. *Semantic Web*, 1(0).
- Chiarcos, C. (2014). Towards interoperable discourse annotation. discourse features in the ontologies of linguistic annotation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC'14*, Reykjavik, Iceland.
- da Cunha, I., Torres-Moreno, J., and Sierra, G. (2011). On the development of the RST spanish treebank. In *Proceedings of the Fifth Linguistic Annotation Workshop, LAW 2011*, pages 1–10, Portland, Oregon.
- Demirşahin, I. and Zeyrek, D. (2014). Annotating discourse connectives in spoken Turkish. In *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*, pages 105–109, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.
- Kallen, J. L. and Kirk, J. M., (2008). *ICE-Guide*. Belfast: Cl6 Ollscoil na Banriona.
- Kallen, J. and Kirk, J., (2012). *SPICE-Ireland: A User's Guide*. Belfast: Cl6 Ollscoil na Banriona.
- Knott, A. and Sanders, T. (1998). The classification of coherence relations and their linguistic markers: An exploration of two languages. *Journal of pragmatics*, 30(2):135–175.
- Prasad, R. and Bunt, H. (2015). From semantic relations in discourse: The current state of iso 24617-8. In *Proceedings of the Eleventh Joint ACL - ISO Workshop on Interoperable Semantic Annotation*.
- Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The penn discourse treebank 2.0. In *Proceedings of the 6 International Conference on Language Resources and Evaluation, LREC'08*. European Language Resources Association.
- Prasad, R., Webber, B., Lee, A., and Joshi, A. (In preparation). Discourse relations in the pdtb 3.0.
- Sanders, T., Spooren, W., and Noordman, L. (1992). Toward a taxonomy of coherence relations. *Discourse Processes*, 15:1–35.
- Sanders, T., Vis, K., and Broeder, D. (2012). Project notes on the dutch project discan. In *Eighth Joint ACL - ISO Workshop on Interoperable Semantic Annotation*.
- Sanders, T., Demberg, V., Evers-Vermeul, J., Hoek, J., Scholman, M., and Zufferey, S. (In preparation). Unifying dimensions in discourse annotation.
- Scholman, M. C., Evers-Vermeul, J., and Sanders, T. J. (2016). A step-wise approach to discourse annotation: Towards a reliable categorisation of coherence relations. *Dialogue and Discourse*, 2(7).
- Stoyanchev, S. and Bangalore, S. (2015). Discourse in customer care dialogues. Poster presented at the DiSpoL 2015 workshop “Identification and Annotation of Discourse Relations in Spoken Language”, October 2015 in Saarbrücken, Germany.
- Sweetser, E. (1990). *From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure*. Cambridge University Press.
- Tonelli, S., Prasad, R., and Joshi, A. (2010). Annotation of discourse relations for conversational spoken dialogs. In *Proceedings of the 7 International Conference on Language Resources and Evaluation, LREC'10*.
- Webber, B. (2009). Genre distinctions for discourse in the penn treebank. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics, ACL'09*.