

## Signalling Implicit Relations: A PDTB – RST Comparison

**Lucie Poláková**

POLAKOVA@UFAL.MFF.CUNI.CZ

**Jiří Mírovský**

MIROVSKY@UFAL.MFF.CUNI.CZ

**Pavína Synková**

SYNKOVA@UFAL.MFF.CUNI.CZ

*Institute of Formal and Applied Linguistics  
Faculty of Mathematics and Physics  
Charles University in Prague  
Malostranské náměstí 25  
118 00 Prague, Czech Republic*

**Editor:** Amanda Stent

Submitted 05/2016; Accepted 07/2017; Published online 12/2017

### Abstract

Describing implicit phenomena in discourse is known to be a problematic task, from both theoretical and empirical perspectives. The present article contributes to this topic by a novel comparative analysis of two prominent annotation approaches to discourse relations (coherence relations) that were carried out on the same texts. We compare the annotation of implicit relations in the Penn Discourse Treebank 2.0, i.e. discourse relations not signalled by an explicit discourse connective, to the recently released analysis of signals of rhetorical relations in the RST Signalling Corpus (RST-SC). The intersection of corresponding pairs of relations is rather a small one, but it shows a clear tendency: unlike the overall signal distribution in the RST-SC, more than half of the signals in the studied intersection are of semantic type, formed mostly by loosely defined lexical chains. Our data transformation allows for a simultaneous depiction and detailed study of the two resources.

**Keywords:** discourse coherence, rhetorical relations, implicit discourse relations, signals, Rhetorical Structure Theory, Penn Discourse Treebank, RST Signalling Corpus

### 1. Introduction and Motivation

In recent discourse-oriented research, great attention has been paid to discourse markers or discourse connectives<sup>1</sup>, which are agreed to be the most apparent anchors of discourse relations, and in this way substantially contribute to discourse coherence. Nowadays, there are even large efforts to thoroughly describe such discourse-relational devices in form of lexicons (e.g. (Stede and Umbach, 1998; Roze et al., 2012)) and other inventories that would serve both linguistic purposes and NLP.

A natural step in the research on discourse coherence is then to answer the general question of how coherence is established if such a connective device is not present between given text segments. From the linguistic viewpoint, we may want to describe what other overtly present language

---

1. disregarding the strictness / looseness of possible definitions of these categories

elements, even elements not directly associated with discourse coherence functions, can play a role in our understanding of such a connection. From the cognitive viewpoint, we may be interested in the way our mind processes such a connection (let us say, an implicit relation), why we all (mostly) understand and interpret a given text in the same way and what part of the overall meaning is inferred from a co-textual, situational or world knowledge context. Finally, from the perspective of automatic text processing, we may want to model discourse coherence with the help of detecting the same signals/features a human normally uses for full comprehension of texts.

In all known attempts, the annotation of implicit relations (and connectives) has been a difficult task. The inter-annotator agreement figures for implicit relations from various discourse annotation projects are mostly calculated together with the figures for explicit relations and thus the actual statistics on implicit relations remains unpublished<sup>2</sup>. Personal discussions with annotators experienced with this task shows that the agreement on implicit relations is perceived rather low, far beyond satisfactory. This was also the reason for postponing such a task in our own work (annotation of discourse phenomena in the Prague Dependency Treebank) for later phases, when the annotators get more experienced in recognizing discourse relations and also when we have gathered feedback from the results of similar projects.

The aim of this article is to contribute to our understanding of how coherence is established in the absence of connective devices, by performing a comparative corpus analysis. In particular, we make use of the existence of multiple annotations for the same Wall Street Journal texts. We focus on the Penn Discourse Treebank 2.0 (PDTB, Prasad et al. (2008b)) annotation of implicit discourse relations, look for their counterparts in the RST Signalling Corpus (RST-SC, Das et al. (2015)) and analyze coherence signals assigned to these relations in the RST signalling annotation. To put it differently, we are interested less in what connective could be inferred to anchor such an implicit connection, and more in what language element(s) in the text (and possibly other pragmatic settings) influenced the PDTB annotator to insert a specific connective token (and no other) and to annotate a specific label (and no other). Our hypothesis is that there must be evidence for the annotator's decision in the text itself and the RST annotation of signals can provide this evidence. Also, it can provide essential empirical information on how far such a description can actually reliably proceed, as we assume that the way humans speak and write is to a certain degree vague.

The results of our study can hopefully be of use for various NLP tasks concerned with discourse phenomena. Finding out how (and to what degree) implicit relations are anchored on the surface can contribute to enhancing systems using discourse-aware features. In shallow discourse parsing and especially classification of implicit relations, there has been noticeable progress in the last years (e.g. (Pitler et al., 2009; Lin et al., 2009; Park and Cardie, 2012) and others), also thanks to the intensive search for the best feature combinations. The latest work in this field (Li and Nenkova, 2014) contains a discussion on sparsity issues of lexical features and offers a comparison of different settings of lexical and syntactic features. The authors arrive at the observation that those lexical

---

2. To our knowledge, there are only a few published inter-annotator agreement figures solely for implicit relations. In the Penn Discourse Treebank 2.0, the percentual agreement between two annotators on setting the extent of the argument spans was 85.1% with an exact-match metric, and 92.6% with a partial match metric (Miltsakaki et al., 2004; Prasad et al., 2008b). Agreement on the inserted connectives was 72% (for 5 semantic categories), (Miltsakaki et al., 2004, p. 7). A recent measurement for implicit relations in the Turkish Discourse Treebank (Zeyrek and Webber, 2008) reports chance-corrected Kappa values of 0.52 for the class level, 0.43 for the type level and 0.34 for the subtype level (Zeyrek et al., 2015).

features computed as most significant for the system performance captured semantic information quite related to the nature of the relation, but also, that these features were mostly domain-specific<sup>3</sup>.

Last but not least, as already addressed in Poláková (2015, pp. 144–154), the present analysis is intended as a springboard for designing an annotation scenario for “discourse relations with no connective” in a future release of the Prague Dependency Treebank (a treebank of written journalistic Czech). Since there already is (apart from explicit discourse relations) a complex annotation of underlying syntax, information structure (topic-focus articulation), genres, pronominal and nominal textual coreference and also bridging anaphora (Bejček et al., 2013), it makes methodologically more sense to focus on already available signals rather than, for instance, on inserting a connective into a connection with an already annotated strong contrastive bridging link<sup>4</sup>.

### 1.1 From Implicit Relations to Signals

In this article, we use the term *implicit discourse relations* in agreement with the Penn Discourse Treebank terminology; this term signifies discourse relations that are not signalled by any connective device in the text (on the surface). More details on the annotation of implicit relations in the PDTB 2.0 are given in Section 2.1<sup>5</sup>.

- (1) [Several organizations, including the Industrial Biotechnical Association and the Pharmaceutical Manufacturers Association, have asked the White House and Justice Department to name candidates (*for judges*) with both patent and scientific backgrounds.] [The associations would like the court to include between three and six judges with specialized training.]

In Example 1 from the PDTB, the two sentences represent two discourse units (discourse arguments, in the PDTB terminology) related to each other with an implicit discourse relation in the PDTB and with a rhetorical relation in the RST-Discourse Treebank (RST-DT, Carlson et al. (2002))<sup>6</sup>. The example represents an exact match in two ways: (i) the text spans of both discourse units match exactly in the two frameworks and (ii) there exists a coherence relation between these units in both annotations. They also represent a partial match<sup>7</sup> in the categories assigned to the relation in both annotations: The sense of the relation in the PDTB was annotated as EXPANSION–RESTATEMENT–SPECIFICATION and the implicit connective “specifically” was inserted, in the RST-DT, the relation was annotated as ELABORATION–ADDITIONAL (the left argument represents the nucleus). Finally, the RST Signalling Corpus provides the relation with two signals:

1. combined; (semantic+syntactic); (repetition + subject\_NP); comment: associations
2. single; semantic; lexical\_chain; comment: A few lexical chains.

These signals of implicit relations, their types, combinations and distribution, and the linguistic implications of their analysis, are the very focus of our research.

3. like e.g. the words *share* or *cent* for Contrast in the financial domain

4. For details on annotation categories regarding phenomena “beyond the sentence boundary” in the Prague Dependency Treebank see Zikánová et al. (2015), Chapters 2–5.

5. To avoid confusion, it has to be stressed again that the term *implicit relations* refers to the implicitness (absence) of a connective device, not to the implicitness of one of the connected discourse units, as in [I THINK THAT] *He has already left because his car is gone.*

6. In the RST analysis, the first sentence consists of two elementary discourse units (EDUs), the first one being split by an embedded unit. The rhetorical relation in question thus holds between a subtree of an RST tree with three terminal nodes on one side and a single terminal node on the other. Compare also Figure 1.

7. For details on label mapping between the two frameworks see Section 3.4.

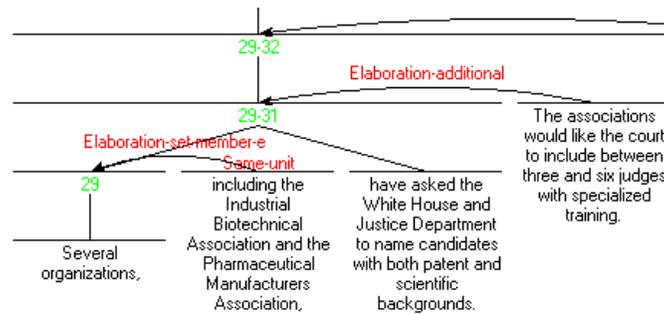


Figure 1: RST analysis (a subtree from the RST Tool) for the sentences in Example 1

## 2. Global and Local Coherence Modeling in Corpora

The leading frameworks in discourse analysis access discourse phenomena from two main perspectives on discourse coherence modeling. The so-called “global”, top-down approaches represent a whole document as a single connected structure (such approaches are also referred to as “deep” discourse parsing), while “local”, bottom-up approaches access discourse phenomena from the syntactic perspective (“shallow” discourse parsing). The most influential frameworks among the former are Rhetorical Structure Theory (RST, Mann and Thompson (1988)), the Segmented Discourse Representation Theory (SDRT, Asher and Lascarides (2003)) and the Discourse Graphbank (Wolf and Gibson, 2005). The latter, “local” direction of discourse analysis is best represented by the lexically grounded approach of the Penn Discourse Treebank (Prasad et al., 2008b), which accesses discourse relations in the first place by searching for their lexical anchors – discourse connectives. Also, the PDTB does not make any claims about the shape of the overall discourse structure.

We are well aware of this basic difference in the theoretical aiming and description methods of the two (three) corpora targeted in this research, namely the Penn Discourse Treebank, (the RST Discourse Treebank) and the RST Signalling Corpus. (A detailed description of these projects follows in Sections 2.1, 2.2 and 2.3.) The different basic departure points of these resources can be observed already in the first step of such an analysis: on the segmentation of a discourse (text) to (elementary) discourse units<sup>8</sup>. The differences in the conception of discourse units and their hierarchization in both approaches are addressed below in the section on argument mapping (3.3). Also, the relations between (among) discourse units are approached differently in these frameworks – whereas the RST investigates *rhetorical relations*, the PDTB uses the term *discourse relations*. Acknowledging the differences between the notions behind these terms, for the purposes of this article we treat these relations in the same way and use the term *discourse relations* or, a more neutral term *coherence relations*.

Nevertheless, both recent developments in the discourse-oriented research community – which is, among other things, comparing the existing frameworks for discourse annotation or mapping them onto one another (e.g. (Versley and Gastel, 2013; Prasad and Bunt, 2015; Scholman et al., 2016)) – and our own experience with discourse annotation (Poláková et al., 2013) indicate that even such theoretical and descriptive variation offers areas of intersection where certain phenomena can be observed with methodological soundness.

8. The RST framework uses the term *discourse units* (DUs), while the PDTB segments are called *discourse arguments*.

## 2.1 Penn Discourse Treebank

The Penn Discourse Treebank (Prasad et al., 2008a) contains annotation of discourse relations over the 1 million word Wall Street Journal corpus in its current version PDTB 2.0 (Prasad et al., 2008b), version 3.0 is forthcoming. The PDTB approach to discourse relations is based on the identification of discourse connectives as anchors of local discourse relations. Following this lexically grounded approach, the annotation in the PDTB first consisted in marking relations signalled by explicit discourse connectives (according to a predefined list of these expressions). The location and the extent of the arguments of explicit relations were not restricted but all relations were assumed to hold between two and only two arguments. As a second step, discourse relations that have no connective device on the surface were annotated. The annotators were instructed to read adjacent sentences within a paragraph and, for each pair of sentences not already linked by an explicit connective, they made a decision as to whether a discourse relation is present. Then they inserted a connective conveying the best the meaning of the relation and provided a label for this relation. Relations with an inserted connective are called implicit<sup>9</sup>. Apart from adjacent sentences within a paragraph, implicit relations were also annotated intra-sententially between clauses delimited by a semi-colon or a colon.

However, in three situations a connective could not be inserted between the adjacent sentences. These situations have led to the introduction of three additional categories: AltLex (alternative lexicalization of a connective), EntRel (entity-based relation) and NoRel (no relation). In the case of AltLex, insertion of a connective would lead to redundancy, since the relation was expressed by an expression or phrase not included in the original list of connectives (e.g. *one reason is, that is why, further*, see Prasad et al. (2010)). In the case of EntRel, only an entity-based relation was detected between the given arguments (annotators were not able to insert an appropriate connective); and finally in the case of NoRel, no discourse relation was perceived between the sentences. The distribution of all types of relations (and NoRels) in the PDTB is the following – overall, there are 40 771 relations, 18 459 of them explicit, 16 224 implicit<sup>10</sup>, 624 signalled by AltLexes, 5 210 EntRels and 254 NoRels. Explicit relations thus represent 45.3% of all relations, the implicit ones 39.8% and AltLexes 1.5%.

Sense annotation in the PDTB 2.0 uses a three-level hierarchy containing 4 general classes, 16 sense types and 23 subtypes (30 possible senses if choosing the most detailed label available Prasad et al. (2008b, p. 2965)). Sense annotation was provided for explicit, implicit and AltLex relations only, as EntRels and NoRels do not indicate (according to the PDTB approach) the presence of a discourse relation. This fact represents an important difference compared to the RST approach – relations which are called entity-based in the PDTB are not different from the other discourse (rhetorical) relations according to the RST-DT.

Separately, the PDTB annotates the attribution of each discourse relation and of each of its two arguments. Attribution is a relation between agents who report some content and the reported content and thus, according to the PDTB approach, it is not a discourse relation. Here, again, there is a discrepancy between the two frameworks – attribution relations in PDTB correspond to regular rhetorical relations in the RST-DT. This issue is further addressed in Section 3.3.

9. The annotators could insert one or two connectives for a single implicit relation, and to each of them assign one or two labels. For the present study we only take into account the first label of the first connective.

10. If we consider implicit discourse relations with two inserted connectives as two relations, the total number of implicit relations in the PDTB is 16 224. If we consider them as one, the total number is 16 053. Later on in this paper, we use the latter number.

## 2.2 RST Discourse Treebank

The RST Discourse Treebank (RST-DT, Carlson et al. (2002)) is a language resource annotated for rhetorical relations over 385 Wall Street Journal articles (176 000 words) selected from the Penn Treebank (Marcus et al., 1993). The chosen texts concern a variety of topics and were annotated manually under the Rhetorical Structure Theory framework (Mann and Thompson, 1988). Rhetorical relations are considered to hold between elementary discourse units (and/or segments composed of these units) which often correspond to clauses but are not restricted in this respect; also, the number of discourse units connected by one relation is not restricted. The RST-DT uses a set of 78 rhetorical relations for the annotation (Carlson and Marcu, 2001) and does not provide information about the connective devices. Relations are categorized according to nuclearity status (i.e. the presence of relevant relational content in one or in more discourse units for a given relation). The annotation captures the global structure of each text in the form of a tree diagram with elementary discourse units as its leaves and rhetorical relations as its edges.

## 2.3 RST Signalling Corpus

The RST Signalling Corpus (RST-SC, Das et al. (2015)) adds an annotation of signalling information to each of the rhetorical relations in the RST Discourse Treebank. Overall, the corpus contains signals for 21 400 relations (Das and Taboada, 2015). For each relation, more than one signal can be present. The taxonomy of signals, based on features identified in previous studies and preliminary corpus work (Taboada and Das, 2013), is hierarchically organized in three levels: signal class, signal type and specific signal. Three values for the signal class are *single*, *combined* and *unsure*, meaning that for a given relation either a single signal was found, or the signalling is combined from one independent and one dependent signal, or no signal was found. The class *single* is divided into nine types<sup>11</sup>. A relation in a given context can be signalled by:

1. a discourse marker - an expression like *because, and, now*
2. reference features - personal, demonstrative or comparative reference
3. lexical features - an indicative word or phrase expressing the relation such as *compared with, explaining, that means, the result is that*
4. semantic features - words (not pronouns) or phrases in a mutual semantic relation present in both/several discourse units of the relation; the semantic relation can be strong (synonymy, antonymy, meronymy, repetition, word pairs like *asked – replied*, general words like *matter* or *thing* referring to the previous context) or relatively weak (i.e. a lexical chain like *selling shares – credit – concern – company – holding, confuses – clearer*)
5. morphological features - tense, aspect or mood change between units
6. syntactic features - a relative, infinitival, participial or imperative clause, interrupted matrix clause, a parallel syntactic construction, reported speech, inversion of subject auxiliary, nominal or adjectival modifier

---

11. In what follows, specific signals for each type are presented only broadly, not in detail. For a detailed description with examples see the RST-SC annotation manual (Das and Taboada, 2015).

7. graphical features - colon, semicolon, dash, parentheses, items in sequence
8. genre features - inverted pyramid scheme, newspaper layout, newspaper style attribution or newspaper style definition
9. numerical features - the number of certain objects is represented by a word in one span (e.g. *three, five*) and these entities are then named in the other span

The class *combined* comprises neither all the possible combinations of types from the class *single* nor combinations of all specific signals. The classification is data-driven and contains the following six types:

1. reference + syntactic features combine a reference feature and a subject noun phrase (NP)
2. semantic + syntactic features combine all semantic features and a subject NP
3. lexical + syntactic features combine an indicative word and a present participial clause
4. syntactic + semantic features combine parallel syntactic constructions and lexical chains
5. syntactic + positional features combine a participial clause and a beginning position of this clause
6. graphical + syntactic features combine a comma and a participial clause

For the intended analysis, it is important to note the difference between combined signals and multiple signals in the RST-SC annotation scheme. Combined signals have two parts – one of them is an independent signal, the other depends on the first one (e.g. in a combined *reference + syntactic* signal, the second signal, the subject NP, “is used to specify additional attributes of the first signal” (Das and Taboada, 2015, p. 9)). On the other hand, multiple signalling refers to the possibility of a relation to be signalled by more than one separate signal (single or combined) functioning independently from each other.

The class *unsure* “refers to those cases in which no potential signals were found or were specified” (Das and Taboada, 2015, p. 33).

The distribution of all signals in the RST Signalling Corpus is presented in Table 3 below in Section 4, together with a comparison to signal distributions only for the subset of relations that have PDTB implicit counterparts. As far as we know, the general distribution was not presented by the authors of the corpus (their distributions are related to individual relations only (Das, 2014)), that is why we present the results of our own measurement.

### 3. The Method

The methodological procedure for the comparative corpus analysis consisted of several steps: First, all three corpora were manually inspected in order to get a basic idea about the properties of the data to be compared. For the RST-DT and the RST-SC we used their respective annotation / visualization tools: the RST Tool 3.45 (O’Donnell, 2000) and the UAM Corpus Tool 3.2i (O’Donnell, 2008). For the PDTB annotation, we used exports from the column-transformed data representation and a recently developed PML-based extension of TrEd, a Prague tool for treebank viewing and searching

(Mírovský et al., 2016), see also Section 3.1. In this way, a sample of six WSJ documents was selected according to their different genre characteristics (Webber, 2009) and the number of implicit PDTB relations (129 sentences, 51 implicit relations). A preparatory survey was conducted by hand (18 matches with RST relations detected). Simultaneously, the data annotated in all corpora (Section 3.1) were converted into a common working format and a procedure for automatic argument mapping was designed (Sections 3.2 and 3.3). The manual survey served as a check on the accuracy of the mapping procedure. Next, label alignment was performed on the basis of hand-crafted sense alignment principles (Section 3.4). Finally, the intended comparative analysis could be carried out on the matching subset of data (Section 4).

### 3.1 Data

The PDTB consists of 2 159 files (documents) with 48 338 sentences (the average number of sentences per document being 22.4) and contains 16 053 annotated implicit relations (see footnote 10). 364 out of the 2 159 files (i.e. 16.9%) were also annotated in the RST-DT with RST trees and RST relation types, and in the RST-SC with information on signals<sup>12</sup>. By number of sentences (8 532 out of 48 338), it represents 17.7% of the whole PDTB (the average number of sentences per RST document<sup>13</sup> being 23.4). These 364 documents represent the data we used for the present study.

### 3.2 Data Conversion

For unification of the data of the PDTB and the RST-SC, we used a framework for treebank annotation and data processing that consists of three core components:

1. the Prague Markup Language (PML)<sup>14</sup>, an abstract XML format designed for annotation of linguistic corpora, especially treebanks,
2. TrEd, a highly customizable tree editor (Pajas and Štěpánek, 2008)<sup>15</sup>, which can be used to browse and edit data in the PML format, and
3. the PML-Tree Query (PML-TQ), a powerful system for querying any data encoded in the PML format (Pajas and Štěpánek, 2009)<sup>16</sup>.

Once a treebank is transformed into the PML format, it can be browsed and edited in TrEd, and searched using the PML-TQ – see for example a transformation of the Penn Discourse Treebank to the PML (Mírovský et al., 2016), or a project of harmonizing 36 treebanks into a common data format and annotation scheme HamleDT (Zeman et al., 2014).

For our task, we needed to combine (i) information from the RST-SC (which includes also the original annotations of the RST-DT tree structures and relation types) and (ii) the annotation of implicit relations from the PDTB, in a single data representation. We proceeded in two steps:

First, the data of the RST-SC were transformed to the PML: The RST tree structures along with the relation types were transformed from the Penn bracketing format, and – at the same time

12. The whole RST-DT (as well as the RST-SC) consists of 385 Wall Street Journal articles; 21 of these 385 files were not annotated in the PDTB.

13. counted on the 364 files annotated both in the PDTB and RST-SC

14. For an on-line documentation for the Prague Markup Language, see <http://ufal.mff.cuni.cz/jazz/PML/>.

15. TrEd is written in Perl and can be easily customized for a desired purpose by extensions that are included into the system as modules. It is available from <https://ufal.mff.cuni.cz/tred/> under the GNU General Public Licence.

16. See e.g. Chapter 8 of Zikánová et al. (2015) for a practical introduction to the query language.

Relations	Count
Implicit relations in the PDTB	16 053 (see footnote 10)
Implicit relations in the part of the PDTB that was annotated also in the RST-SC (364 files)	2 892 (18% of all implicit relations in the PDTB)
Imported implicit relations (both arguments match with nodes in the RST trees)	2 286 (79% of all implicit relations in the 364 files)
- without a counterpart RST relation (arguments are not siblings in the RST tree)	1 499 (65.6% of all imported relations)
- with a counterpart RST relation (arguments are siblings in the RST tree)	787 (34.4% of all imported relations, 27.2% of all implicit relations in the 364 files)
Imported implicit relations with a match of the counterpart RST relation (see Section 3.4)	472 (60% of imported relations with a counterpart RST relation, 20.6% of all imported relations, 16.3% of all implicit relations in the 364 files)

Table 1: Overview of implicit relations throughout the transformation process

– the information on signals taken from the XML files was mapped into the tree structures using references to positions in the Penn bracketing format.

Second, the implicit relations from the PDTB were mapped onto the transformed data, whenever the arguments of an implicit relation could be matched with node spans in the RST trees. Matching of the arguments was performed via comparing rawtext representations of the arguments, as they were given in the two source formats (the Penn bracketing format for the RST-SC, and the column-transformed annotation of the PDTB data). Systematic inconsistencies between rawtext representations of arguments in the two sources were taken into account: before comparing the arguments, we removed paragraph separation marks, parentheses, and leading and trailing punctuation. Table 1 gives an overview of numbers of implicit relations in the PDTB and in the RST-SC/PDTB overlapping data. As our study only focuses on implicit phenomena, the explicit PDTB relations and AltLexes and their arguments were not mapped in this stage.

Apart from the data transformation itself, an extension to TrEd for displaying the transformed data was implemented<sup>17</sup>. Figure 2 displays a part of the RST tree structure for the text from Example 1 in TrEd (see also Figure 1 for its depiction in the original RST Tool). In our implementation, the RST tree structure and relation types are displayed along with signals for the RST relations and matching PDTB implicit discourse relations. Red arrows/polylines represent RST relations; orange arrows represent implicit discourse relations from the PDTB. For the time being (and for technical reasons), the information about any relation is at the start node of the relation; the type of the RST relation is depicted in red, signals are depicted in magenta, and comments to signals in brown. Semantic types (senses) of the PDTB implicit relations and inserted connectives are in orange.

17. The extension is freely downloadable from inside TrEd using its extensions management tool. Scripts for transforming the original RST-DT, RST-SC and PDTB data into the PML are a part of the extension; for details, see the dedicated web page: <https://ufal.mff.cuni.cz/rst>.

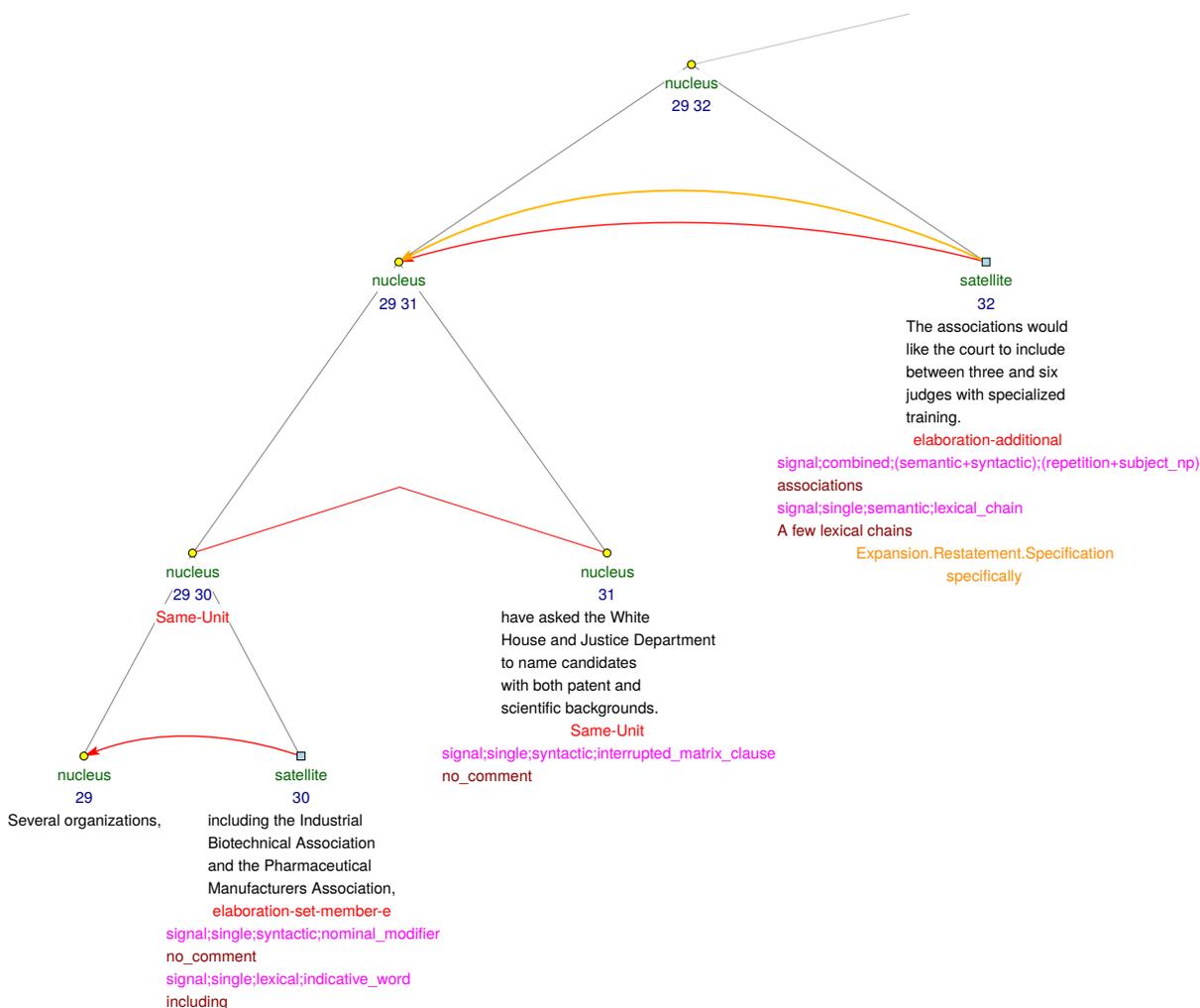


Figure 2: Joint annotation in TrEd for the sentences in Example 1 (see also Figure 1)

### 3.3 Argument Mapping

As already mentioned in Section 2, the different theoretical assumptions behind the two corpora lead to different strategies in the delimitation of discourse units (discourse arguments). Moreover, arguments of implicit relations in the PDTB 2.0 differ slightly from arguments of explicit relations: arguments of implicit relations were partly predefined by the principle to annotate an implicit relation between adjacent sentences within a paragraph (an argument being represented by either the whole sentence or its part)<sup>18</sup>.

18. The adjacency rule for the annotation of arguments of implicit relations was loosened for the later annotation of English biomedical texts in BioDRB (Prasad et al., 2011) by allowing the annotators to search also for a remote left argument of an implicit connective. The authors were able to reduce the percentage of NoRels, i.e. of cases where no relation to the immediately preceding sentence could be found, from 1.15% in the PDTB to 0.9% in the BioDRB (Prasad et al., 2014, p. 924).



PDTB implicit relation	RST relation	Count
Expansion.Conjunction	Elaboration-additional	85
Expansion.Conjunction	List	80
Expansion.List	List	60
Expansion.Restatement.Specification	Elaboration-additional	36
Contingency.Cause.Reason	Explanation-argumentative	26
Comparison.Contrast	Elaboration-additional	20
Contingency.Cause.Reason	Elaboration-additional	17
Expansion.Restatement.Specification	Elaboration-general-specific	17
Contingency.Cause.Result	Elaboration-additional	15
Expansion.Instantiation	Example	13
Comparison.Contrast.Juxtaposition	Contrast	11
Comparison.Contrast	Contrast	10
Expansion.Restatement.Specification	Explanation-argumentative	10
Contingency.Cause.Result	Consequence-s	9
Expansion.Conjunction	Comparison	9
...		

Table 2: Fifteen most frequent correspondence pairs of implicit PDTB and RST relations (787 relations in total); (substantial) semantic mismatches are highlighted in grey

- (2) [“Overall demand still is very respectable,”] [says Christopher C. Cole, group vice president at Cincinnati Milacron Inc., the nation’s largest machine tool producer.] [“The outlook is positive for the intermediate to long term.”]

The results of the argument mapping demonstrate two basic tendencies: if an RST discourse unit matches exactly with a PDTB argument of an implicit relation, the PDTB argument typically consists of more RST elementary discourse units (EDUs) within one subtree. These EDUs mostly correspond to clauses and clause-like syntactic structures within one sentence. On the contrary, non-matching PDTB arguments consist of RST nodes that do not form a subtree. Further, where the discourse units of two similar relations do not match, it is mainly due to the exclusion of a part of the sentence from the PDTB argument – often due to the exclusion of attribution spans (e.g. *He said*) from the PDTB arguments, compare again Example 2 and Figure 3. These near-matches are not included our study; at this point, our aim was to reliably find exactly matching relation pairs, without further intervention. In a less strict approach, some relations with attribution spans could be probably considered matching counterparts if the attribution spans were detected in the data, analyzed separately and removed or disregarded (with possible manual work). This would lead, in our opinion, to an enlargement of the intended dataset.

### 3.4 Semantic Labels

The research in this article does not focus on the correspondence and mapping of labels in the two annotation frameworks – this is current work of e.g. Sanders et al. (submitted). Moreover, for the

given two categorizations of discourse (rhetorical) relations in the PDTB 2.0 and the RST-DT, such a mapping is a difficult task, (not only) since there is a big difference in the granularity of the two taxonomies. The PDTB 2.0 uses a three-level hierarchy of senses with 30 possible end-level values (Prasad et al., 2008b) whereas the RST-DT annotation distinguishes 78 types of relations in 16 classes (Carlson et al., 2001). Nevertheless, for the purposes of this article, it appeared necessary to provide at least partial correspondence links for the labels on both sides. The links are partial in the sense of mapping only those relations that actually occurred in our dataset (not the whole taxonomies)<sup>19</sup>, and also in the sense of capturing the agreement only on a reasonably general semantic level. For the analysis of signals, there has to be agreement in the two corpora that a given signal is actually a signal of one particular, even if coarse-grained, category.

Let us demonstrate this situation on Example 1 above: it can be observed that the PDTB label EXPANSION–RESTATEMENT–SPECIFICATION and the RST label ELABORATION–ADDITIONAL share basic semantic features and thus can be treated as corresponding on a general semantic level: they are both additive, they expand the piece of information given in A (a demand for candidates for judges with specific skills) by adding B (how many judges are required). On the other hand, the subtle semantics of specifying or giving a detail is not encoded in the RST ELABORATION–ADDITIONAL relation. There can be a better fit, the ELABORATION-GENERAL-SPECIFIC relation.

In principle, types of relations that agreed within one of the four first-level PDTB-defined classes were assessed as corresponding, with few exceptions: within the Expansion-like group, substantially different relations were not matched, e.g. PDTB EXPANSION–CONJUNCTION and RST EXAMPLE; within the Contingency-like group, conditionals were not treated as equal to causals; and so on. In other words, we tried to keep the matching strict. Only in such a way can we rely on the basic assumption required for our study: in the cases of relations with corresponding semantic categories, we can consider the signals for the RST relations to be signals for the PDTB implicit relations as well. According to this method, from the 787 relation pairs we had at our disposal based on the argument mapping, 60% (472) relation pairs have similar semantic characteristics and can be further worked with. These 472 relation pairs have altogether 674 signals, i.e. 1.43 signals per relation<sup>20</sup>. Table 2 shows the fifteen most frequent correspondence pairs of implicit PDTB relations and RST relations. Rows highlighted in grey mark pairs of relations whose semantic labeling could not be treated as matching.

#### 4. The Comparison

In the comparative analysis itself, we focus on two main ways of comparison. First, we analyze the distribution of all signal types in the whole studied subset of matching implicit PDTB relations against the distribution of these signal types in the whole RST Signalling Corpus. The matching dataset is quite small compared to the sizes of both source corpora; nevertheless – without aspiring on generalizing our results on the whole treebanks – it allows us to observe linguistically relevant and interesting tendencies for the studied phenomenon (4.1). A more fine-grained analysis concerns the *semantic* type of signal, since it proved to be the most frequent type of signalling in the studied dataset (Section 4.1.1). Second, we analyze signal type distributions with regard to different PDTB

19. See again footnote 9 in Section 2.1.

20. In fact, it was only 471 relations, as one of them (between text spans 164 and 163 in the file wsj\_0681) was not annotated with any signal in the RST-SC.

Single	RST	∩	Combined	RST	∩	Unsure	RST	∩
syntactic	29.8	1.3	semantic+syntactic	7.4	13.5	unsure	5.3	12.2
semantic	24.8	51.2	reference+syntactic	1.9	5.5			
discourse marker	13.3	1.3	syntactic+semantic	1.4	5.5			
lexical	4.9	2.8	graphical+syntactic	0.7	0			
graphical	3.5	1.8	lexical+syntactic	0.4	0			
genre	3.2	0	syntactic+positional	0.2	0			
reference	2	4						
morphological	1.1	0.6						
numerical	0.1	0.3						

Table 3: Distribution of types of signals in the whole RST Signalling Corpus (the RST column) and in the matched 472 relations (674 signals) corresponding to implicit PDTB relations (the ∩ column); all values are in percents.

senses (Section 4.2). As there are not enough occurrences for all PDTB relation sense (sub)types in our dataset, we concentrate on the most frequent ones.

#### 4.1 Overall Signal Distributions

Table 3 presents percentages for occurrences of all signal types in the whole RST Signalling Corpus (385 documents, 21 400 relations, 29 300 signals) and further percentages for a subset of RST relations that have implicit PDTB counterparts agreeing in argument spans and in the (basic) semantic characteristics (472 relations, 674 signals in total).

Considering first only the general RST-SC signal distribution, it can be observed that more than a two-third majority of signals (approx. 68%) is represented by only three categories of signalling: syntactic (29.8%), semantic (24.8%) and discourse markers (13.3%). Discourse markers, although this category is in general perceived broader than the category of discourse connectives, are much less frequent in the RST-SC than are explicit connectives in the PDTB (the proportion of explicit relations in the PDTB is 45.3%, cf. Section 2.1 above)<sup>21</sup>.

In total, 201 types of RST discourse markers were identified (the “type” here refers to a unique string, so e.g. *if* and *only if* are counted as two types). The combination of semantic and syntactic signals, i.e. one of the semantic signal subtypes (repetition, lexical chain, synonymy, meronymy or general word) in combination with a subject NP, is the fourth most common signal (7.4%), followed by a 5.3% of unsure cases (signals not found or non-specified). The proportion of no other signal type (nor their combination) exceeds 5%.

Proportions of the signal types in the subset of RST relations corresponding to implicit PDTB relations show some substantial differences. As expected, the proportion of discourse markers drops dramatically (to 1.3%). A manual inspection of discourse markers in this subset indicates that these

21. We are aware that a direct comparison is not possible here and that this observation is only very rough: first, the RST-SC represents approximately one sixth of the PDTB size; second, the annotation approaches to discourse connectives and discourse markers differ significantly; and third and most importantly, building a hierarchical tree structure for the whole document results in the existence of many more relations per word.

markers are distinct<sup>22</sup> from the PDTB connectives, for example expressions like *now*, *particularly*, *most importantly*, *unfortunately* or *after all*. In this way we can confirm a basic assumption that implicit PDTB relations generally do not correspond to relations with discourse markers in the RST-SC.

If we take a closer look at the *lexical* signal type (indicative words, alternate expressions), we observe that the concept of PDTB AltLexes and that of a lexical signal type in the RST-SC include similar items – e.g. *compared with*, *followed by*, *including*, *reason*, *explain(ing)*, *still*, *like*, *finally*, but also longer strings like *no matter how much* or *in the past two weeks* etc. Since the PDTB annotates AltLexes only under specific conditions and thus many such expressions and phrases were not marked at all (see Section 2.1), it can be assumed that this category would be more numerous in the PDTB. This is not so much the case in the RST-SC: the proportion of the lexical signal type in the whole RST-SC is only 4.9%. In our subset of PTDB – RST matching relation pairs, the proportion slightly drops (to 2.8%). Single lexical signals in the studied subset comprise altogether 19 indicative words, among which there are 4 occurrences of the expression *even*, (as in Example 3); the remaining lexical signals occur only once each. In the PDTB, *even* is treated as a connective modifier (Prasad et al., 2007, p. 9), not as a separate connective. As a result, in Example 3, the PDTB annotator did not annotate an explicit *even*-relation or an AltLex but instead inserted an implicit *also*-connective (and the label EXPANSION–CONJUNCTION). The remaining three occurrences are analogical: the *even* expression<sup>23</sup> does not modify any other connective-like expression, but has a scope over other parts of the sentence.

- (3) [The pilot program was received well (by teachers and students), but there wasn't reason enough to sign up.] [We **even** invited the public to stop by and see the program, but there wasn't much interest.]

In our opinion, this example demonstrates a well known issue of delimitation of the category of discourse connectives, discourse markers and other strong lexical cues of discourse connections, may they be called alternative lexicalizations, lexical signals or secondary connectives (as in Rysová and Rysová, 2014).

The most apparent distribution change between the two datasets is represented by the drop in the most frequent signal type – the *syntactic* signal type (from almost 30% to 1.3%). In the implicit subset, there are only 9 cases, all of them represented by parallel syntactic constructions, cf. Example 4.

- (4) “How the hell can you live with yourself?” he erupts at a politician. [“You twist people’s trust.”] [You built your career on prejudice and hate.]

Signal: single; syntactic; parallel\_syntactic\_constructions; comment: you twist ~ you built

22. apart from one annotation error

23. a focusing particle or a rhematizer in the Prague approach to information structure and a possible discourse connective in the analysis of discourse relations

The very low proportion of syntactic signals in the studied subset is most likely caused by the fact that the PDTB implicit relations are in the vast majority realized inter-sententially, whereas most of the RST syntactic signal subtypes apply only intra-sententially. The parallelism of syntactic constructions, as demonstrated in Example 4, is one of the few possible syntactic signals applying also between individual sentences. It appears that syntactic signals, on their own the strongest signalling cue in the RST-SC, can play only a restricted role as sole signals of implicit PDTB relations. On the other hand, the proportions of three types of combined signals with a syntactic component shows a visible increase in the studied subset (first three cells in the “Combined” column in Table 3): *semantic + syntactic* from 7.4 to 13.5%, *reference + syntactic* from 1.9 to 5.5% and *syntactic + semantic* from 1.4 to 5.5%. These signals are represented mostly by the following subtypes, respectively: *lexical chain + subject NP* and *repetition + subject NP*, *personal reference + subject NP*, *parallel synt. constructions + lexical chain*. In these combinations, again, only the *parallel syntactic constructions* syntactic subtype applies; the *subject NP* component does not function as an independent signal, but is always dependent on the previous component in a combination (Section 2.3).

Quite in the opposite direction from syntactic signalling goes the *semantic* type of signalling, it increases by 26.4% to more than a half (51.2%) of all signals in our subset. This fact, in our opinion, is the most expressive evidence about the nature of signalling of the PDTB implicit relations. That is why we analyze the semantic type of signals individually, in Section 4.1.1 below.

Finally, a great distribution difference can be detected in the *unsure* category, it increases from 5.3% to 12.2% in the studied subset. This is also the only tag that never co-occurs with other types of signals, which means that the number of unsure tags in our subset (82) is also the number of relations with this (and no other) labeling.

#### 4.1.1 SEMANTIC SIGNALS, LEXICAL CHAINS

According to the RST-SC annotation manual, a semantic signal, unlike most other signals, “has two components (words or phrases), each belonging to one of the spans. The components are in a semantic relationship with each other, such as synonymy, antonymy, and lexical chain...” (Das and Taboada, 2015, p. 20). It has six subtypes: synonymy (e.g. *Scandinavian Airlines System ~ SAS*), antonymy (*profit ~ loss*), meronymy, i.e. set–member relation (*computer firms ~ Microsoft Corp.*), repetition (*Avis ~ Avis*), indicative word pair, i.e. very closely semantically related words or phrases (*asked ~ replied; resigned ~ succeeded*) and lexical chain (*personal computers ~ desk-top computers, microprocessors, minicomputers, mainframes*). A lexical chain is defined as follows:

“Words or phrases in the respective spans are identical or semantically related. Unlike the repetition feature, words or phrases in lexical chains do not refer to object or entity, but they belong to the class of indefinite or common nouns and also other syntactic categories, such as adjectives, verbs and adverbs. Lexical chain differs from synonymy, antonymy, meronymy and indicative word pair in a significant way. While in the latter categories, the semantic relation between the words or phrases is very strong and can be specified, words or phrases present in a lexical chain are related to each other by a relatively weak semantic connection.” (Das and Taboada, 2015, p. 21)

There are altogether 473 *semantic* signal occurrences in the studied dataset (single and in combination). 161 are *single.semantic* signals appearing as a sole signal of a relation, i.e. the relation does not have combined or multiple signalling<sup>24</sup>. Interestingly, all these semantic signals are of the sub-

24. Especially here, note the difference between multiple and combined signals, as explained above in Section 2.3.

type *lexical chain*; that means that none of the remaining 5 categories function as a sole signal of an implicit PDTB relation. The remaining 312 cases with semantic signals are represented by 60 other combinations of (single, combined or both) signals, with most of the combinations not exceeding 1% of the occurrences of all cases with semantic signals<sup>25</sup>. Although it is difficult to further analyze such a variety of cases, one tendency is traceable: also here, with only one exception, at least one signal from each of the 60 combinations is represented by the *lexical chain* subtype.

We have therefore further briefly analyzed the comments for the *lexical chain* subtype. There are altogether 408 occurrences of this subtype in our dataset (or 86% of all semantic signals). The comments for *lexical chain* typically specify the exact wording of the lexical chains in question, cf. Example 5 annotated with the RST label EXPLANATION-ARGUMENTATIVE and the following signal:

- (5) [Excluding the gain, P&G’s earnings were close to analysts’ predictions of about \$1.40 a share for the quarter.] [Wall Street had expected a modest rise in the company’s domestic sales and earnings, and more substantial increases in overseas results.]

Signal: single; semantic; lexical\_chain, comment: (close to analysts’ predictions) ~ (expected a modest rise)

On the other hand, in 247 cases the annotators only indicated that there are more lexical chains, cf. Example 1 above, or they did not provide any comment at all (40 cases). Sometimes, the comment included also a note that the lexical chain is a rather loose one.

If we sum up our observations so far and relate them to the number of PDTB relations in the studied subset (472), the following can be stated: 161 relations are signalled by a single semantic signal of the subtype *lexical chain*, which is a rather weak semantic connection. A further 82 relations are assigned the *unsure* tag; one relation has no signal assigned. That implies that more than half (exactly 51.7%) of all relations in our intersection are signalled weakly or their signals are difficult to detect.

## 4.2 Signals across Implicit Relations

There are only 17 distinct PDTB relation types (or 3rd-level subtypes) in the studied subset, and their distribution is quite uneven. For instance, none of the TEMPORAL class relations exceeds 10 occurrences and, on the contrary, we can observe a clear prevalence of relations from the EXPANSION class (355 out of 472, or 75%). This is partly influenced by the nature of the studied data and would be different for other genres and registers, and partly, as we believe, by the implicit nature of the studied relations. Because of the low number of instances we present results only for eight most frequent relation (sub)types from three class level categories (EXPANSION, CONTINGENCY and COMPARISON). For each such relation, we present two most common signals. The results are summarized in Table 4<sup>26</sup>.

The signal distribution across the (relatively) frequent relations confirms at first sight our findings for the overall signal distribution: the leading signal for five of these relations is *single; semantic* or

25. They appear either as combined signals (semantic + other type) or as multiple, independent signals of one relation (single semantic + single semantic; single semantic + combined; more than two signal types etc.).

26. As an attachment to this article, a larger table with full signal distributions across the mapped PDTB relations can be found at <https://ufal.mff.cuni.cz/rst>.

*lexical\_chain*, followed by *unsure* as the leading signal for the remaining three relations. If we cluster the relations according to their general semantic class, it appears that EXPANSION class relations are signalled more by lexical chains whereas the relations from CONTINGENCY and COMPARISON classes were more likely to have *unsure* signalling. Altogether, no other striking differences in signal distributions can be observed for individual relation types.

The semantic weakness of the *semantic.lexical\_chain* signal subtype and the frequency of the *unsure* signals go hand in hand with other facts about semantic weakness from the PDTB annotation: (i) EXPANSION–CONJUNCTION is known to be a relatively broadly and loosely defined relation in the PDTB 2.0 annotation (as opposed to the restriction of its definition for the upcoming PDTB 3.0 relation taxonomy, cf. Prasad et al. (in preparation)) and (ii) there is no subtype-level label annotated for most Contrast relations in our subset<sup>27</sup>, see Table 4. All these partial observations suggest that the relations in the studied subset, regardless of how we defined them, were in a large number difficult to treat for the annotators of both systems. In other words, despite the apparatus of a complex and detailed data-driven signal taxonomy, the cues leading to the interpretation of these particular relations were weak and could not be identified precisely.

## 5. Discussion

The signalling annotation of the RST-SC has a high information value for the presented analysis and possibly also for the purposes of discourse parsing and other automatic text processing tasks. It provides direct access to signals of discourse coherence other than the most apparent cues in the connections regarded as implicit. We are now able to study discourse functions of expressions such as *more* or *differently*, quantify the role of syntactic features and coreference (at least for our limited sample of texts), etc. We can also look at possible combinations of various types of signals. And, we also learn where the presented types of discourse annotation have their limits: it is where semantics comes into play.

From the perspective of Prague multilayer-annotated treebanks, it can be observed that *semantic* signals in the RST-SC terminology partly correspond to Prague *bridging anaphora* annotations. On one hand, we are convinced that a finer subclassification of semantic signals is possible – compare the types of bridging links and further proposed subtypes in Chapter 4 in Zikánová et al. (2015). On the other hand, from its nature, a semantic signal is in principle not a surface signal, but requires the inclusion of inferential processes. It is only natural that a high fraction of the semantic signals in implicit relations, namely those annotated as *lexical chains*, are very difficult to assess, and for the annotators (or any two interpretators) to agree on. Also the Prague annotation of bridging anaphora, although further classified and provided with extensive annotation guidelines, has the lowest agreement figures from all the annotated phenomena annotated in the Prague Dependency Treebank (Zikánová et al., 2015, p.96) (compare also the discussion on p. 60–62 on co-hyponymy, definition of common world knowledge and the necessity of inclusion of Word-Net-like databases in similar tasks).

Our study leads to an indirect observation that the two discourse annotation frameworks are consistent with each other in pointing out places with weak signalling or no signalling of coherence at all. To put it simply, implicit relations are indeed implicit, not signalled, or signalled vaguely. We believe that at this point of linguistic description, with such a high degree of semantically anchored

---

27. There was the option for the PDTB annotators to provide only a higher-level sense label for a relation if they could not decide about the lower-level label.

PDTB implicit relation	Count	Signals	%
Expansion.Conjunction	165	signal; single; semantic; lexical_chain signal; combined; (semantic+syntactic); (lexical_chain+subject_np)	46.6 13
Expansion.Restatement.Specification	82	signal; single; semantic; lexical_chain signal; combined; (semantic+syntactic); (lexical_chain+subject_np)	43.6 12.1
Expansion.List	63	signal; single; semantic; lexical_chain signal;combined;(syntactic+semantic); (parallel_synt._constr.+lexical_chain)	59.2 18.3
Contingency.Cause.Reason	34	signal; unsure signal; single; semantic; lexical_chain	50 47.1
Contingency.Cause.Result	37	signal; unsure signal; single; semantic; lexical_chain	40.5 37.8
Expansion.Instantiation	33	signal; single; semantic; lexical_chain signal; unsure	54 12
Comparison.Contrast	19	signal; unsure signal; single; semantic; lexical_chain	47.6 28.6
Comparison.Contrast.Juxtaposition	11	signal; single; semantic; lexical_chain signal; unsure	57.1 14.3

Table 4: Eight most frequent PDTB implicit relations in the studied subset (472 relations) and two most frequent signals for each (with relative frequencies within signals for the given implicit relation)

coherence relations, we have reached the borderline of what information corpus annotation can provide. Trying to annotate semantic phenomena beyond this point, in our opinion, only results in getting unreliable and inconsistent data with very limited use for NLP purposes. The only possible way for us is to accept a certain degree of vagueness and perhaps to learn to detect places in texts with weak coherence or vagueness instead of imposing a certain type of connective or even discourse meaning on them.

From another perspective, the interpretation of implicit connections can be further studied using the manual Czech translations of the WSJ-part of the PDTB collected in the Prague Czech–English Dependency Treebank (PCEDT 2.0, Hajič et al. (2011)). It offers an ideal opportunity to look for possible explicitations of implicit relations by Czech connectives or other surface cues.

## 6. Conclusions

The study presented in this article made use of a portion of the English Wall Street Journal texts having been annotated for discourse phenomena from two different theoretical perspectives, namely the Penn Discourse Treebank 2.0 and the RST Signalling Corpus. Despite the theoretical and annotation differences (global vs. local approach to text analysis, different segmentation strategies, different

sets of labels for coherence relations etc.), we have shown that there is a common denominator for a comparison of the two annotations. We have been interested in answering the question of how implicit PDTB relations are signalled by means of the RST-SC signal annotation. We have arrived at several observations we believe can be of use for any discourse researcher concerned with implicit discourse phenomena or comparing discourse annotation schemes in general.

(i) A matching intersection of the annotations, in terms of a discourse (rhetorical) relation and the two units it relates, does exist in the compared datasets. We took into account relation pairs with at least broadly corresponding semantic labeling but we aimed for exact argument matching; the resulting intersection is therefore not very large. It comprises 472 implicit PDTB relations (out of 2 892 implicit relations in the part of the PDTB also annotated in the RST-SC) signalled by 674 RST-SC signals<sup>28</sup>. That is why we do not aspire to generalize our results for the whole treebanks. Nevertheless, we believe our results allow us to observe relevant and interesting tendencies for the studied phenomena.

(ii) The comparative analysis showed that to a large extent (51.2%), implicit PDTB relations are signalled by signals of semantic nature; these signals are anchored in the semantics of specific lexical chains in the arguments. Lexical chains are characterized as a rather weak type of semantic connection by the authors of the RST-SC themselves. These lexical chains are either specified word for word in the comments in the annotations, or, in many cases, the annotators indicated that there can be more lexical chains taking part in interpreting the relation, and did not specify them. Further, in 19%, semantic signals appeared in combination with syntactic types of signals (subject NP, parallel syntactic constructions). In 12.2%, the signalling of implicit PDTB relations was unsure, compared to the 5.3% of unsure cases in the whole RST-SC. If we take the number of PDTB implicit relations in the studied dataset into account, slightly more than a half of them (51.7%) are signalled either by a sole signal of the type *single;semantic;lexical chain* or are unsure. These observations indicate that annotation of implicit relations, at least in the studied subset, cannot be easily solved by looking for overtly present signals. Their nature is to a large extent semantic and, moreover, often outside the scope of well definable semantic categories (synonymy, antonymy, set–member relationship, etc.). Our analysis therefore seems to confirm the fact that annotation/classification of implicit relations is a challenging task both for humans and for automatic methods in NLP applications.

(iii) Last but not least, for the purposes of this study the data of the two treebanks were transformed to a common format (PML) and a unified visualization in the TrEd tool was developed. One of its extensions, the PML-Tree Query, also enables for the first time to search the imported treebanks for various phenomena at once.

## Acknowledgements

This work has been supported by the projects GA CR 17-06123S of the Czech Science Foundation and Multilingual Corpus Annotation as a Support for Language Technologies (LH14011) of the Ministry of Education, Youth and Sports of the Czech Republic. The work has been using language resources and tools developed and/or stored by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (LM2015071). We would like to thank the two anonymous reviewers for their helpful suggestions.

---

28. The intersection would be probably larger if different treatment of attribution was taken care of, and, naturally, it would be larger if also explicit PDTB relations were taken into account.

## References

- Nicholas Asher and Alex Lascarides. *Logics of Conversation*. Cambridge University Press, 2003.
- Eduard Bejček, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Magda Ševčíková, Jan Štěpánek, and Šárka Zikánová. *Prague Dependency Treebank 3.0*. Univerzita Karlova v Praze, MFF, ÚFAL, Prague, Czech Republic, 2013.
- Lynn Carlson and Daniel Marcu. Discourse tagging reference manual. Technical Report ISI-TR-545, Los Angeles, 2001. Available from <<http://www.isi.edu/~marcu/discourse/tagging-ref-manual.pdf>>.
- Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, pages 1–10, 2001.
- Lynn Carlson, Mary Ellen Okurowski, Daniel Marcu, et al. *RST Discourse Treebank*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, 2002.
- Debopam Das. *Signalling of Coherence Relations in Discourse*. PhD thesis, Simon Fraser University, Burnaby, Canada, 2014.
- Debopam Das and Maite Taboada. *RST Signalling Corpus Annotation Manual*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, 2015.
- Debopam Das, Maite Taboada, and Paul McFetridge. *RST Signalling Corpus*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, 2015.
- Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Uřešová, and Zdeněk Žabokrtský. *Prague Czech-English Dependency Treebank 2.0*. Institute of Formal and Applied Linguistics, Prague, Czech Republic, 2011.
- Junyi Jessie Li and Ani Nenkova. Reducing Sparsity Improves the Recognition of Implicit Discourse Relations. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 199–207, Philadelphia, PA, USA, 2014.
- Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 343–351, 2009.
- William C. Mann and Sandra A. Thompson. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8:243–281, 1988.
- Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

- Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. Annotating discourse connectives and their arguments. *Proceedings of the HLT/NAACL Workshop on Frontiers in Corpus Annotation*, pages 9–16, 2004.
- Jiří Mírovský, Lucie Poláková, and Jan Štěpánek. Searching in the Penn Discourse Treebank using the PML-Tree query. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 1762–1769, 2016.
- Michael O’Donnell. RSTTool 2.4: a markup tool for Rhetorical Structure Theory. In *Proceedings of the International Conference on Natural Language Generation*, pages 253–256, 2000. Available from <<http://www.wagsoft.com/RSTTool/>>.
- Michael O’Donnell. The UAM Corpus Tool: Software for corpus annotation and exploration. In *Proceedings of the XXVI Congreso de AESLA, Almeria, Spain*, pages 3–5, 2008. Available from <<http://www.corpustool.com>>.
- Petr Pajas and Jan Štěpánek. Recent advances in a feature-rich framework for treebank annotation. In *Proceedings of the International Conference on Computational Linguistics*, pages 673–680, 2008.
- Petr Pajas and Jan Štěpánek. System for querying syntactically annotated corpora. In *Proceedings of the Joint Conference of the Annual Meeting of the ACL and the International Joint Conference on Natural Language Processing*, pages 33–36, 2009.
- Joonsuk Park and Claire Cardie. Improving implicit discourse relation recognition through feature set optimization. In *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 108–112, 2012.
- Emily Pitler, Annie Louis, and Ani Nenkova. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the Annual Meeting of the ACL and the International Joint Conference on Natural Language Processing*, pages 683–691, 2009.
- Lucie Poláková. *Discourse Relations in Czech*. PhD thesis, Charles University in Prague, Prague, 2015.
- Lucie Poláková, Jiří Mírovský, Anna Nedoluzhko, Pavlína Jínová, Šárka Zikánová, and Eva Hajičová. Introducing the Prague Discourse Treebank 1.0. In *Proceedings of the International Joint Conference on Natural Language Processing*, pages 91–99, 2013.
- Rashmi Prasad and Harry Bunt. Semantic relations in discourse: The current state of ISO 24617-8. In *Proceedings of the Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)*, pages 80–92, 2015.
- Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie Webber. *The Penn Discourse Treebank 2.0 Annotation Manual*. Philadelphia, 2007.
- Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. *Penn Discourse Treebank Version 2.0*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, 2008a.

- Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse Treebank 2.0. In *Proceedings of LREC 2008*, pages 2961–2968, 2008b.
- Rashmi Prasad, Aravind Joshi, and Bonnie Webber. Realization of discourse relations by other means: Alternative lexicalizations. In *Proceedings of the International Conference on Computational Linguistics*, pages 1023–1031, 2010.
- Rashmi Prasad, Susan McRoy, Nadya Frid, Aravind Joshi, and Hong Yu. The biomedical discourse relation bank. *BMC Bioinformatics*, 12:188–205, 2011.
- Rashmi Prasad, Bonnie Webber, and Aravind Joshi. Reflections on the Penn Discourse Treebank, comparable corpora, and complementary annotation. *Computational Linguistics*, 40:921–950, 2014.
- Rashmi Prasad, Bonnie Webber, Alan Lee, and Aravind Joshi. *Discourse Relations in the PDTB 3.0*. in preparation.
- Charlotte Roze, Laurence Danlos, and Philippe Muller. LEXCONN: a French lexicon of discourse connectives. *Discours. Revue de linguistique, psycholinguistique et informatique*, (10), 2012. doi: 10.4000/discours.8645.
- Magdaléna Rysová and Kateřina Rysová. The centre and periphery of discourse connectives. In *Proceedings of the Pacific Asia Conference on Language, Information and Computing*, pages 452–459, 2014.
- Ted Sanders, Vera Demberg, Jet Hoek, Merel Scholman, Fatemeh Torabi Asr, Sandrine Zufferey, and Jacqueline Evers-Vermuel. Unifying dimensions in discourse relations: How various annotation frameworks are related. *Corpus Linguistics and Linguistic Theory*, submitted.
- Merel Scholman, Jacqueline Evers-Vermeul, and Ted Sanders. Categories of coherence relations in discourse annotation: Towards a reliable categorization of coherence relations. *Dialogue & Discourse*, 7(2):1–28, 2016.
- Manfred Stede and Carla Umbach. DiMLex: A lexicon of discourse markers for text generation and understanding. In *Proceedings of the International Conference on Computational Linguistics*, pages 1238–1242, 1998.
- Maite Taboada and Debopam Das. Annotation upon annotation: Adding signalling information to a corpus of discourse relations. *Dialogue and Discourse*, 4:249–281, 2013.
- Yannick Versley and Anna Gastel. Linguistic tests for discourse relations in the Tüba-D/Z corpus of written German. *Dialogue and Discourse*, 2:142–173, 2013.
- Bonnie Webber. Genre distinctions for discourse in the Penn Treebank. In *Proceedings of the Joint Conference of the Annual Meeting of the ACL and the International Joint Conference on Natural Language Processing of the AFNLP*, pages 674–682, 2009.
- Florian Wolf and Edward Gibson. Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2):249–287, 2005.

- Daniel Zeman, Ondřej Dušek, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zdeněk Žabokrtský, and Jan Hajič. HamleDT: Harmonized multi-language dependency treebank. *Language Resources and Evaluation*, 48(4):601–637, 2014.
- Deniz Zeyrek and Bonnie Webber. A discourse resource for Turkish: Annotating discourse connectives in the METU corpus. In *Proceedings of the Workshop on Asian Language Resources at the International Joint Conference on Natural Language Processing*, pages 65–71, 2008.
- Deniz Zeyrek, Işın Demirşahin, A. B. S. Çallı, and Murathan Kurfalı. Annotating implicit discourse relations in Turkish: The challenge of corrective discourse relations. In *Abstracts of the International Pragmatics Association (IPRA) Conference*, pages 455–456, 2015.
- Šárka Zikánová, Eva Hajičová, Barbora Hladká, Pavlína Jínová, Jiří Mírovský, Anna Nedoluzhko, Lucie Poláková, Kateřina Rysová, Magdaléna Rysová, and Jan Václ. *Discourse and Coherence. From the Sentence Structure to Relations in Text*. Studies in Computational and Theoretical Linguistics. ÚFAL, Praha, Czechia, 2015. Available from <[https://ufal.mff.cuni.cz/books/preview/2015-zikanova\\_full.pdf](https://ufal.mff.cuni.cz/books/preview/2015-zikanova_full.pdf)>.