Charles University in Prague
Faculty of Mathematics and Physics

Doctoral Thesis

DISCOURSE RELATIONS IN CZECH

Mgr. Lucie Poláková

Prague, 2015

Doctoral Thesis

Mgr. Lucie Poláková

Supervisor of the doctoral thesis:
Prof. PhDr. Eva Hajičová, DrSc.

Discourse Relations in Czech

Study programme: Computer Science
Specialization: Mathematical Linguistics

ÚSTAV FORMÁLNÍ
A APLIKOVANÉ LINGVISTIKY

Prague, 2015

I declare that I carried out this doctoral thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

Prague, June 29, 2015. Lucie Poláková

Název práce: Diskurzní vztahy v češtině
Autor: Mgr. Lucie Poláková
Ústav: Ústav formální a aplikované lingvistiky
Vedoucí práce: Prof. PhDr. Eva Hajičová, DrSc.,
Ústav formální a aplikované lingvistiky
Klíčová slova: koherence, diskurzní vztahy, diskurzní konektory, jazykový korpus, anotace

Abstrakt: Tato doktorská práce se zabývá lingvistickou analýzou diskurzních vztahů jakožto jednoho z aspektů textové koherence. Diskurzními vztahy rozumíme významové vztahy mezi jednotlivými propozicemi v textu, tzv. diskurzními argumenty. Cílem práce je ucelený popis diskurzních vztahů v češtině a jeho vtělení do anotačního schématu Pražského závislostního korpusu. Práce je rozdělena do tří částí: První z nich je zaměřena na teoretický popis diskurzních vztahů a rozbor vhodnosti různých metodologických postupů při korpusovém zpracování. Druhá část podrobně popisuje navržené schéma pro anotaci diskurzních vztahů a proces vzniku takto značeného korpusu včetně evaluace konzistence značených dat. V poslední části práce se pak věnujeme některým problematickým okruhům při užití navrženého schématu a jejich řešení.

Title: Discourse Relations in Czech
Author: Mgr. Lucie Poláková
Department: Institute of Formal and Applied Linguistics
Supervisor: Prof. PhDr. Eva Hajičová, DrSc.,
Institute of Formal and Applied Linguistics
Keywords: coherence, discourse relations, discourse connectives, language corpus, annotation

Abstract: This doctoral thesis is devoted to linguistic analysis of discourse relations as one of the aspects of discourse coherence. Discourse relations are semantic relations holding between propositions in a discourse (discourse arguments). The aim of the thesis is a complex description of discourse relations in Czech and its application in the annotation scheme of the Prague Dependency Treebank. The thesis is divided into three parts: The first one is focused on the theoretical description of discourse relations and on analysis of adequacy of various methodological concepts in corpus processing. The second part describes in detail the proposed scheme for the annotation of discourse relations and the process of the corpus build-up including the evaluation of consistency of the annotated data. Finally, in the last part of the thesis, we address some problematic issues arisen with the employment of the proposed scheme and we look for their possible solutions.

Contents

Acknowledgement

I would like to express my gratitude to the Institute of Formal and Applied Linguistics for the excellent support, both professional and personal. It is my colleagues, many of whom I may call my friends, who motivated me and kept encouraging me throughout the years of my PhD studies.

Most importantly, I would like to thank to Prof. Eva Hajičová for her insightful advice and for her unfailing readiness to address any of my questions and concerns. She also deserves thanks for helping me, among other students of hers, to get in contact with numerous international scientists and projects in our field.

I'm very grateful to my team mates for all the fruitful discussions, reflection on my ideas and for a huge source of inspiration by our common work on corpus projects. Pavlína Jínová, my annotator twin, offered her linguistic talents whenever I needed feedback; Jiří Mírovský is to be thanked in particular for teaching me advanced corpus querying and for many technical and practical comments to my analyses. Veronika Pavlíková and Magdaléna Rysová are to be praised for their tireless annotation effort. A big thank belongs to Šárka Zikánová who stood by my side throughout the years both professionally and personally, teaching me how to organize my academic and family responsibilities. For further suggestions and general support, I would like to thank Vladimír Patera, Eduard Bejček, Markéta Lopatková, Silvie Cinková, Zuzanna Bedřichová, Jana Šindlerová, Zdeněk Žabokrtský, Jaroslava Hlaváčová, Hana Major Sládková, Lenka Sándor and Michal Kebrt, among many others.

A word of gratitude belongs also to Prof. Aravind Joshi and his influential research team at the University of Pennsylvania, and to Yannick Versley at the University in Heidelberg who both kindly accepted me at their research institutes and made my stays there very inspiring and pleasant.

Last but not least, this thesis would never have come into being without the support of my family and in particular of my wonderful parents.

ACKNOWLEDGEMENT

This dissertation thesis is based on projects investigated at the Institute of Formal and Applied Linguistics, Charles University in Prague. It was supported by the Grant Agency of the Czech Republic (project P406/12/0658) and by the Ministry of Education of the Czech Republic (project LH14011). It has been using language resources and tools developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education of the Czech Republic (project LM2010013).

1

Introduction

1.1 Discourse Relations

It is widely acknowledged that discourse (text) is more than a sequence of individual pieces of information, more than a succession of utterances. A well-formed discourse, spoken or written, is a coherent, meaningful whole, and its individual segments are interconnected by a number of diverse relations.

In this dissertation thesis, we investigate one such aspect of discourse coherence: "discourse relations" or, in other words, semantic relations that connect discourse segments. This is illustrated by Example (1)[1] which presents a connection of three such discourse segments[2].

(1) *In a Delhi hotel, the Beatles registered under the name of Brown and party. Reporters later found a Sikh sitar player giving lessons to George Harrison, while John Lennon was found trying to play a snake-charmer's flute.*

One discourse relation (signaled by the expression *later*) occurs between the two sentences and expresses temporal successiveness; another (signaled by the expression *while*) occurs intra-sententially, within the second sentence, and indicates simultaneity.

We propose a framework for a systematic description of discourse relations and their ways of expression in Czech, based on observations on written contemporary texts, predominantly from journalistic domain. Methodologically, we put special emphasis on corpus processing, formal representation and applicability of the analysis in natural language processing (NLP). There are two points of departure for our research: the framework for the description of discourse relations in the Penn Discourse Treebank 2.0 (PDTB) for English, and the tectogrammatical

[1] from the BBC Magazine

[2] There might be more discourse segments in Example (1), depending on their definition.

(deep syntactic and semantic) representation of Czech sentences in the Prague Dependency Treebank (PDT).

In our study, we first focus on conceptual issues and possibilities of a linguistic analysis on such a complex level of language description. Then, we propose a description scheme for discourse relations in Czech and carry out an extensive analysis on authentic texts. The analysis goes hand in hand with our aim to design a first corpus of Czech annotated for discourse relations (among other discourse-related phenomena, like coreference and associative anaphora). The procedure of the corpus creation is described. In the last part of the thesis, we discuss the outcomes and contributions of the annotation project, addressing in particular two topics: (i) mismatches between analyses of sentence structure and discourse structure and (ii) consequences of presence or absence of discourse connectives and other discourse-structuring devices in discourse relations.

1.2 Motivation

Coherence and discourse structure are nowadays burning research topics in international linguistics. The most elaborated concepts, however, are developed predominantly for English. English was also the first language to be annotated for discourse-related phenomena (Carlson et al. 2002). This fact literally invites the speakers of other languages to test and verify such theories on other languages, possibly typologically very different.

It is not that the Czech linguistics would lack discourse-oriented studies. The discipline of text linguistics has been forming since 1970's in Czechoslovakia, cf. Section 2.3.1. But, so far, neither a Czech (Czechoslovak) nor a foreign approach to discourse phenomena has been tested systematically and on a large scale on Czech language material.

These facts open an excellent opportunity to attempt at a systematic description of discourse phenomena in Czech using authentic texts and methods of corpus linguistics. The idea suggests itself even more as we already have at our disposal complex corpus analyses of sentential phenomena (the annotation layers of the Prague Dependency Treebank, cf. Section 2.2.3).

Also, the topic of this thesis has the advantage to build on the results of our diploma thesis (Mladová 2008a, in Czech) where we targeted the meeting points of sentential syntax (and its representation in the Prague treebanks) and relations in discourse, as a preparatory study for creation of a future annotation scheme for discourse.

1.3 Objectives

The starting point for the research in this thesis, and also a common denominator for all related theories and corpus projects, is the basic question: **What makes a discourse/text coherent?** What are the means in a natural language for connecting pieces of information together and so enabling successful communication? Can coherence of a discourse be modeled, can it be formalized in any way? And to what extent?

The main, overall aim of this thesis is to contribute to the general knowledge about discourse coherence and to find, at least partial, answers to these questions for Czech. In particular, the research objectives of the thesis are the following:

- to address the conceptual question of a linguistically adequate description for discourse based on existing frameworks

- to introduce the points of departures for our approach to discourse (The Praguian Functional Generative Description and its application in Prague corpora and the Penn Discourse Treebank project)

- to define, delimit and classify discourse relations in Czech, in particular for the purposes of their formalized representation in a language corpus; and, in this way, to contribute significantly to a design of the discourse layer of language description in the PDT

- to document thoroughly the process of creation of the Prague Discourse Treebank 1.0

- to perform a quantitative analysis of the annotation results and put it into context of the theoretical frame used

- to detect problematic issues in the annotation, interpret them linguistically and propose a possible solution for their further treatment

The nature of the thesis objectives brings along some limitations which we are well aware of: First, it is the restrictions we face when working mostly with texts from a single domain. The language data we use are fairly big (almost 50,000 sentences), taking into consideration what is achievable by the costly manual annotation, but they are not representative. Therefore we sometimes use also other corpora to support our claims with evidence from other domains. Still, most of our findings are necessarily influenced by the convention of language use in journalism.

Next, the nature of the task of applying a theoretical concept to a large amount of language data has the consequence that not all phenomena that deserve attention can be addressed in detail. This is why, once having the annotated treebank at our disposal, we had to select only a few subtopics for an in-depth linguistic analysis. Our selection of these subtopics was motivated by (i) repeatedly occurring inconsistencies in the annotations and by (ii) weak points discovered in the theoretical concept.

Finally, the task of a large corpus creation is always by its nature a collective work. Even the linguistic concept as an annotation base can hardly be a deed of a single person (not to mention data management, hundreds of hours of annotation work, revisions etc.). In this respect, we want to clarify our own contribution to the linguistic work behind the corpus creation and to the development of the corpus itself. The initial idea to draw inspiration from the Penn Discourse Treebank research team comes from Eva Hajičová who also supervises the whole project. The role of the author of this thesis in the project was first to develop an annotation scheme for discourse relations in Czech and then to lead and coordinate the annotation process (while being one of the annotators). Some basic decisions on how the Czech annotation scheme for discourse relations should be shaped were discussed with Eva Hajičová, Šárka Zikánová and Zuzanna Bedřichová. Later, during the intensive annotation period, the theoretical issues and annotation feedback were mostly discussed with Pavlína Jínová and Jiří Mírovský. Pavlína became the most active and experienced annotator, and Jiří was responsible for the annotation tool and data management. The research reported in this thesis is therefore necessarily influenced by and dependent on the work of these fellow researchers. We greatly acknowledge all this work and the support of all the team members during the time of creating this thesis.

1.4 Structure of the Thesis

The thesis has three major parts: the theoretical part (Chapter 2) is devoted to the notion of discourse relations in general; Chapter 3 describes the linguistic scheme proposed for annotation of discourse relations in the Prague Discourse Treebank 1.0 (PDiT 1.0) and the process of the annotation. In Chapter 4, we report on the annotation results and analyze the annotated material in various aspects.

In **Chapter 2**, we first define the basic terms used in our research (Section 2.1). In Section 2.2, we introduce the theoretical background, methodology and describe the Prague Dependency Treebank as the main resource of Czech material analyzed. Section 2.3 reviews Czech and international linguistic approaches to analyzing discourse structure and coherence and offers an overview of the state-of-the-art discourse annotation projects. In Section 2.4, we address general theoretical and methodological issues in discourse analysis stemming from the frameworks introduced earlier; we describe basic linguistic properties of discourse relations and discuss their possible ways of treatment in the intended corpus-oriented analysis.

In **Chapter 3**, we first outline the two fundamental decisions underlying the creation of the Prague Discourse Treebank – inspiration by the Penn Discourse Treebank and annotating discourse relations on top of syntactic trees (Section 3.1). Section 3.2 describes the technicalities: the data representation and format, the annotation tool and the interface for treebank querying. Section 3.3 takes a closer look at the annotation procedure itself. In Sections 3.4 to 3.7, a detailed characteristics of the annotation of discourse connectives, discourse arguments, discourse relations and other discourse-related phenomena is presented. Section 3.8 gives an overview of post-annotation checking procedures and offers an evaluation of annotation consistency.

Chapter 4 is divided into three parts: Section 4.1 offers a detailed corpus statistics for all annotated phenomena. In Section 4.2, we analyze in detail one of the most apparent source of annotation inconsistencies – the places where the syntactic structure and the discourse structure clash, mainly in terms of discourse unit delimitation and location. Finally, Section 4.3 explores a possible analysis extension on the basis of our experience so far: the topic of implicit discourse relations.

Chapter 5 summarizes the findings of the previous chapters and the contributions of the thesis.

2

Theory and Methodology

2.1 Key Terms

In this section, we provide definitions of the basic notions used in this thesis. Their detailed characteristics follows later on, in Chapter 3.

2.1.1 Text and Discourse

The terms of *text* and *discourse* are broad concepts that have undergone a diverse development in different times, areas and scientific disciplines. To avoid terminology confusion, we offer here a brief overview of the development of both terms in text-/discourse-oriented linguistic research and specify the way these terms will be further used in this thesis.

European linguistic schools in general prefer to use the term *text* to *discourse*, following W. Dressler's concept of *Textlinguistik* established as a discipline in the seventies of the last century (Dressler 1972; de Beaugrande and Dressler 1981). So does the British Hallidayan school (Halliday and Hasan 1976). The Czech linguistic tradition follows the European one; the field of our interest is established in the Czech linguistic community primarily as *textová lingvistika* (text linguistics).

On the other hand, American and America-inspired linguistic schools work predominantly with the term *discourse*. In general, in English written linguistic research, we speak nowadays about the "structure of sentence and discourse", about "discourse and dialog", about "spoken and written discourses". The term *discourse* is problematic in Czech partly because of its various translation possibilities connected to other meanings (*diskurz, text, promluva, jazykový projev*), partly due to its generally ambiguous nature. *Discourse* was until recently mostly associated with the interdisciplinary approach of CDA – critical discourse analysis (e.g. Fairclough 1989) or even as a term from other disciplines, like sociology and literary science (Foucault, Bachtin etc.). In linguistics, it is mostly related to stylistics.

There are various studies distinguishing the properties of *texts* from the properties of *discourses* in different approaches (cf. Schiffrin 1994, p. 21 or Tárnyiková 2002, p. 19). With the exception of the mentioned geographical difference, the notions of *text* and *discourse* vary in different approaches in many aspects: *text* is sometimes treated as written communication, an artefact whereas *discourse* is spoken communication. Also, *text* can be treated as a static concept (a product) whereas *discourse* is a dynamic concept (the process of text creation).

In this thesis, the diverse interpretations of both terms are disregarded and both terms are used as synonyms, with the preference of the term *discourse*. One of the main sources of inspiration comes from the University of Pennsylvania and the Penn *Discourse* Treebank project (Prasad et al. 2008), so adopting its English terminology has appeared not only as a convenient, but even a necessary decision.

Discourse is in our approach understood as a written or spoken form of communication, as a unit of communication which consists of one or more utterances, it is coherent and comprehensible. When we speak about *text*, we mainly refer to the actual corpus texts we conduct our research on. Also, we use the well-established complex expressions like *discourse relations*, *discourse structure*, *discourse units* and *discourse connectives* on one hand, but we do not avoid expressions like *text segment*, *textual coreference* and similar expressions.

2.1.2 Coherence and Cohesion

The terms *coherence* and *cohesion* of a discourse are often used inconsistently in linguistic literature. In some approaches, the two terms are used as synonyms (e.g. Hrbáček 1994, p. 9). But mostly, *coherence* refers to semantic interconnectedness of a text whereas *cohesion* is the demonstration of coherence on the surface, at the level of language expressions (Daneš et al. 1987, p. 633). The terminological heterogeneity of these terms is thoroughly discussed e.g. by Hoffmannová (1993).

In this work, we accept the latter view: coherence is in our approach understood as the semantic interconnectedness and consistency of a discourse. In terms of reception, coherence is a necessary prerequisite for the recipient's ability to assign meaning (intended by the author) to a sequence of text units.

2.1.3 Aspects of Discourse Coherence

There are many factors that participate in creating discourse coherence. According to Czech grammar books (cf. Daneš et al. 1987, p. 685), each discourse unit contains at least one element that connects it with the surrounding discourse units. The Hallidayan school, for instance, distinguishes five aspects (conjunctions, reference, substitution, ellipsis, lexical cohesion) that together organize a text as a neatly woven "texture" (Halliday and Hasan 1976, p. 2). Ten years later, Grosz and Sidner speak about three structures in a discourse that are in mutual relationship: linguistic structure, intentional structure, attentional state (Grosz and Sidner 1986, p. 177).

It is beyond the scope of this thesis to offer a detailed account of all aspects of discourse coherence. We have drafted the possible ways of discourse analysis earlier (cf. Mladová 2008a, pp. 26–28) and in terms of corpus annotation, we have listed the different "layers" of discourse analysis that are now available in our 2014 study (Poláková 2014, pp. 246–248)[1]. However, having pointed out the variety of ways of discourse analysis should put the role of analysis of discourse relations into context of the general account of discourse coherence.

2.1.4 Discourse Relations

The expression *discourse relations* has two interpretations. The broader one, where *discourse* is roughly equal to *text* (as explained above), refers to all relations that can be found in a discourse, including coreferential and associative relations (bridging), thematic structure etc.

In its narrower sense, the term *discourse relations* covers only such type of coherence relations that express a semantic connection between two discourse segments, often anchored by an explicit operator (a discourse connective or some alternative of it).[2]

[1] The most prominent analyses concern: referential structure, associative links (bridging), discourse relations, rhetorical structure, temporal structure, intentional structure, thematic structure, graphical and phonological structure. We also put emphasis on the fact that determining the so-called "pragmatic" aspects of discourse analysis is not easy, mainly because there are multiple views on what this domain actually includes: it can be intentional structure of a discourse, communicative functions, speech act analysis, the so-called pragmatic discourse relations, subjectivity, inferences, presuppositions etc.

[2] Even here, the terminological diversity of the subject is high. In different approaches, these relations are called: coherence relations (e.g. Hobbs 1979; Kehler 2002), rhetorical relations (Rhetorical Structure Theory (Mann and Thompson 1988), Segmented Discourse Representation Theory (Asher and Lascarides 2003)), conjunctive relations (Martin 1992; Stede 2008), informational coherence relations (Wolf and Gibson 2005), discourse relations (Miltsakaki et al. 2004) and so on. In the Czech terminology, for instance, the terms used are "mezivýpovědní vztahy obsahově sémantické" or "vztahy rematické" (Hrbáček 1994, p. 52).

In accordance with the Penn Discourse Treebank terminology, we use this term in the narrower sense. For the broader sense, to avoid ambiguity, we prefer to use the terms *coherence relations* or *relations in a discourse.*

2.1.5 Discourse Connectives

Language expressions whose function is to connect pieces of text into a meaningful whole are called *discourse connectives* (henceforth also DCs)[3]. This category includes devices operating both between sentences and within them, cf. *later* and *while* in Example (1) above for the two respective cases. Also here, following the PDTB, we define a discourse connective as a predicate of a binary relation that takes two discourse units (mainly clauses or sentences[4]) as its arguments, cf. Webber et al. (1999). A DC combines these units to larger ones, signaling a semantic relation between them. In the Prague annotation scenario, most of the connectives are morphologically inflexible and they usually do not act as grammatical constituents of a sentence. Like sentence modality markers, they are "above" or "outside" the proposition. DCs are represented by coordinating conjunctions (e.g. *and, but*), some subordinators (e.g. *because, if, while*), some particles (e.g. *also, only*) and sentence adverbials (e.g. *afterwards*), and marginally also by some other parts-of-speech – mainly in case of fixed compound connectives like *in other words* or *on the contrary.*

2.1.6 Discourse Arguments

The two discourse units building a discourse relation are referred to as *discourse arguments.* Semantically, they are text spans expressing a certain proposition[5] (an action, a state, an event, etc.). Asher (1993) calls them *abstract objects* and offers a detailed classification of them (Asher 1993, p. 1).

Syntactically, a discourse argument can be any possible representation of an abstract object. The most typical discourse argument is a single clause with a finite verb; it may be also a connection of clauses, a (compound) sentence, but also participial and infinitive constructions and nominalizations.

[3] Other terms are e.g.: *discourse cues, cue phrases, discourse markers* etc. The term *discourse markers* is, nevertheless, in our approach a wider concept: we treat discourse connectives as a subset of discourse markers.

[4] Throughout this thesis, a *clause* denotes a structure with a single predication ("věta jednoduchá") whereas *sentence* is understood as a structure "from full stop to full stop", consisting of one or more clauses.

[5] We further use the term *proposition* in a linguistic sense for an elementary predicate structure in a natural language, not in sense of propositional logic.

2.1.7 Semantic Types

For the semantic categories of discourse relations, we use the term *semantic types.* This differs from the PDTB terminology where the term *discourse senses* is used. In the present thesis, we only speak about *senses* when referring to the PDTB annotation scheme and categories. We haven't adopted the PDTB term *sense* in this particular case because it is used with a different interpretation in the FGD tradition, cf. Section 2.2.2.

Other terms used in this thesis, mainly concerning the tectogrammatical analysis of the Prague Dependency Treebank, are introduced gradually, as they appear in the text.

2.1.8 Typographical Conventions

Throughout the thesis, we use the following typographical conventions: Examples with no annotation are printed in italics. In annotated examples, Argument 1 of a discourse relation is printed in italics, Argument 2 is in bold.[6] The discourse connective (or its alternative) is underlined, cf. Example (2).

(2) *Průpravu jsem měl všeho druhu.* Třeba **při rozvozu jsem denně přenesl pěkných pár tun na zádech.**

I had training of all kinds. For example **during deliveries, I daily moved the weight of a good few tons on my back.**

Unless stated otherwise, the examples come from the Prague Dependency Treebank. Artificially constructed examples are marked "LP". The English translations of authentic Czech examples are often influenced by the translation limits. Due to the language differences, some of the translations are the nearest possible approximations to the Czech originals. Consequently, we do not use literal translations of the Czech examples; we only provide literal translations for every case where it is crucial to the understanding.

[6] Be aware of the different strategy in naming the arguments in different corpora, cf. Section 3.5.3.

2.2 Approaches, Methods and Data

In this section, we introduce the basic features of the theoretical and methodological background of our research and characterize the data we carry out our analysis on. Specific starting points for our research within these frameworks are discussed separately in Section 3.1.

2.2.1 Functional Generative Description

The Functional Generative Description (FGD) is a formal framework for natural language description proposed in the 1960's by Petr Sgall and further developed by him and his group (Sgall 1967, Sgall et al. 1969, 1986, Panevová 1980). It is based on the Prague functional and structural linguistic tradition. Its main features are the **stratificational approach** to language description, the use of **dependency syntax** with the notion of **verb's valency** and the inclusion of the **description of the information structure** (topic-focus articulation) into the analysis of a sentence. In the FGD framework, the center of the sentence is the predicate verb; other sentence constituents including the subject are directly or indirectly dependent on it. In accordance with the stratificational approach in FGD, there are several levels of language description, the lowest one being the level of form, corresponding to the surface manifestation of the sentence, and the highest one corresponding to the level of linguistic meaning. The units of neighboring levels are in the relation of form and function: a linguistic form on a lower level of description represents a specific function of a higher level.

In the application of FGD for the scenario of the Prague treebanks, there are four layers, a layer of the tokenized text (w-layer – word layer) and three annotation layers. The lowest annotation layer is the morphological layer (m-layer), the next one, analytical layer (a-layer) corresponds to surface syntax, and the highest level of annotation, which represents underlying syntactic structure and semantic relations in a sentence, is called the tectogrammatical layer (t-layer). The a-layer and the t-layer contain records of sentences as tree structures. The annotation planes are interconnected; from higher ones there are links to lower ones and, at the same time, a piece of information assigned "lower" is projected up to the higher layer. So, in a tectogrammatical tree structure, information on morphological categories assigned to a certain language unit on the morphological layer is retained.

2.2.2 The Concept of Discourse in the FGD Framework

Based on the previous section, we now define the concept of discourse in the Functional Generative Description. Discourse in FGD is **the use of a language as a system in the process of communication**; a discourse is thus understood as a sequence of utterances ("sled výpovědních událostí"). The sense ("smysl") of an utterance consists of the meaning ("význam") of the sentence as a language unit with a specification of the reference of all its referring units (Sgall et al. 1986, p. 17). This implies that discourse relations cannot be understood as some "grammar of the text", it is rather a broader term: discourse is an interconnected network of syntactic, semantic and pragmatic relations.

A question arises whether to speak in the FGD about another level of language description, the level of discourse. It would build a superstructure above the underlying syntactic (tectogrammatical) level. The concept of forms and functions in stratificational approach of FGD would certainly enable the understanding of a clause (a form at the syntactic level of description) as a unit, whose function is reflected at the level of discourse. The levels of description would thus remain intertwined. What is problematic, however, is that a correct interpretation of discourse or even of its individual parts needs more information than a detailed description and linking language levels. It is necessary to take into account extra-linguistic reality, not only the language context, but also the situational one. A "step aside" from the stratification system of language description needs to be made. One should start with the assumption that certain relations in discourse are close to the systematic language description and there are certain patterns and regularities to be investigated. The final understanding of a discourse is, nevertheless, dependent on its anchoring in the communication process, which itself is unique and unrepeatable.

The representation of discourse-related phenomena in Prague treebanks is an extension of the tectogrammatical (underlying syntactic) sentence level. It is neither a separate level "above" the sentential one nor does it belong to it. It is a solely practical decision to annotate discourse (and for that matter, coreference and bridging relations) on the tectogrammatical tree structures. We should keep in mind that theoretically we want to hold apart the underlying representation of a sentence (i.e. tectogrammatics = syntactico-semantic analysis including topic-focus articulation) and the phenomena "beyond the sentence boundary". And, from the top-down perspective, we do not want to build a new, uniform layer of systematic language description for discourse, since, as already stated, discourse-level phenomena stand "a step aside" from the stratal language system.

2.2.3 The Prague Dependency Treebank

The Prague Dependency Treebank (PDT) is a project of the team of researchers from the Institute of Formal and Applied Linguistics (ÚFAL), initiated in 1995. It is a language resource of Czech journalistic texts (two dailies, an economic weekly and a weekly scientific journal[7]) containing approx. 2 million tokens taken over from the Czech National Corpus (CNC). The PDT texts are provided with complex and interlinked annotations of morphology, surface syntax, underlying syntax and semantics (approx. 50 000 sentences, i.e. 0.8 million tokens are annotated on all layers), but also some coreference relations, information structure of a sentence, annotation of named entities etc. are represented. The treebank, or more precisely the 0.8 million tokens annotated on three levels so far, are being continuously updated and enriched by manual annotations of different linguistic phenomena.

For the purposes of this work, we further describe the basic features of the tectogrammatical representation (TR).

2.2.4 Tectogrammatical Representation

In the PDT, the underlying sentence structure and semantics is represented on the so-called tectogrammatical level by a tree structure, cf. Figure 2.1. A tectogrammatical tree consists of tectogrammatical nodes (t-nodes) and edges. The t-nodes represent content words; function words (prepositions, auxiliary verbs, etc.) are represented as attribute values. Also, t-nodes can be newly added to the structure, e.g. a t-node representing a pro-dropped subject or other elided element in reconstructed elliptical constructions. The edges between t-nodes typically express dependency, i.e. they represent the relation between a governing and a dependent node. The semantic type of this relation is a property (attribute) of the edge. However, in the tree representation, it is reflected as one of the most important attributes of the dependent node. This attribute is called the tectogrammatical *functor* (syntactico-semantic label). There are also auxiliary, non-dependency edges in the tectogrammatical representation: they indicate other types of relation – coordinate structures, other specific syntactic functions or they are of technical nature.

[7] (Lidové Noviny, Mladá fronta Dnes, Českomoravský profit, Vesmír)

The tectogrammatical tree structures capture the following aspects of sentences:

- syntactic and semantic dependencies
- syntactic/lexical derivation (t-lemmas)
- fine-grained morphological information
- coordination, apposition, parenthesis
- valency
- information structure (topic-focus articulation)
- grammatical and textual coreference
- ellipsis restoration

According to the concept of valency in FGD, the verb complements can be divided into two groups: actants (or inner participants) and free modifications. Actants are assigned according to the valency frame of the verb; unless coordinated, each actant can appear in the valency frame of the given verb only once. There are five types of actants, represented by the following functors: ACT – actor, PAT – patient, EFF – effect, ADDR – addressee, ORIG – origin. The semantic scale of the free modifications is wide (functors of time, space, direction, manner, causality, etc.). More free modifications of the same type can be assigned to a single verb without coordination.

Further, there are several subgroups of tectogrammatical functors particularly important for the purposes of this work. They are the functors for the meanings of coordinate structures (thus, not describing a type of syntactic dependency but a relation between coordinated items), the functor for conjunction modifiers (CM), functors for expressing attitude of the author towards the content (ATT), functor for modal characteristics (MOD) and the functor for reference to the preceding context (PREC).[8]

Figure 2.1 presents a tectogrammatical tree structure for the sentence in Example (3). The technical root of the tree is displayed in the upper left corner. The effective root of the tree is the conjunction *a* (*and*) connecting (with non-dependency edges) two main clauses governed by the verbs *myslit* (*to think*) and *těšit se* (*to look forward*). Their common (pro-dropped) subject *já* (*I*) is represented by a single generated node with the t-lemma substitute #PersPron. The dependent clause introduced by the verb *být* (*to be*) is connected to its governing clause by a dependency edge with the functor EFF.

[8] For a detailed classification of functors cf. Mikulová et al. 2006.

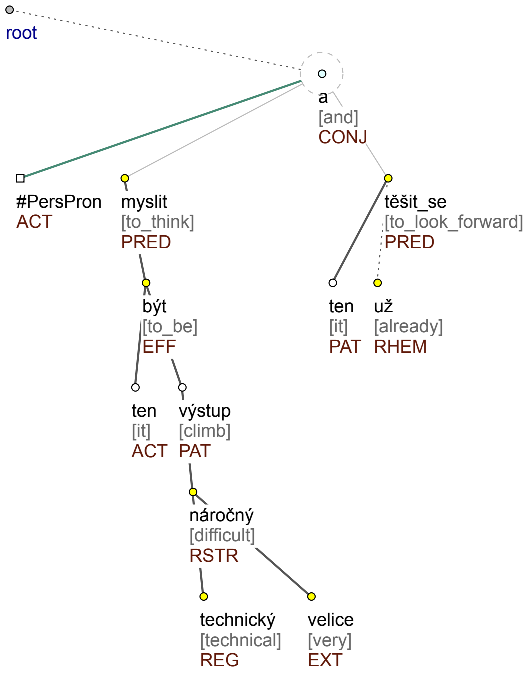


Figure 2.1: A tectogrammatical tree structure

(3) *Myslím, že je to velice technicky náročný výstup, a už se na to těším.*

I think it is technically a very difficult climb, and I'm already looking forward to it.

In our discourse-oriented research, the tectogrammatical analysis is particularly useful for its concept of syntactic dependency and coordination. Also, the semantico-syntactic labeling and certain nodes representing connecting expressions and are potentially relevant for analysis of discourse relations. The advantages of the tectogrammatical analysis for analysis of discourse are discussed thoroughly in Section 3.1.2.

2.2.5 Data versions: The Prague Discourse Treebank 1.0 and The Prague Dependency Treebank 3.0

The Prague Discourse Treebank 1.0 is a result of a subproject of the research group concerned with linguistic phenomena "beyond the sentence boundary". It contains the same texts as the PDT. For the PDiT 1.0, the existing tectogrammatical annotations of the version PDT 2.5 (Bejček et al. 2011) were taken as a base. A new annotation layer, portraying (i) discourse relations, their connectives, arguments and semantic types and (ii) relations of extended textual coreference and bridging relations were added, resulting in The Prague Discourse Treebank 1.0. It was released in November 2012 (Poláková et al. 2012c, Poláková et al. 2013). Later on, the discourse-related annotations were updated and extended for the PDT 3.0 release, which is, up to the present, the newest version publicly available (Bejček et al. 2013, Mikulová et al. 2013)[9]. Both projects have also detailed web documentation.[10] Table 2.1 sums up and compares the discourse-related annotations present in the two treebank versions.

The work reported in the present thesis is a result of research spanning over several years. In our research, we first worked with the unpublished annotated material, then with The Prague Discourse Treebank (PDiT 1.0) and finally with the most updated version, the PDT 3.0. This is why we mostly report numbers for both treebanks, if both were available at the time of the research.

[9] Both treebanks can be downloaded from the LINDAT-Clarin Repository: http://hdl.handle.net/11858/00-097C-0000-0008-E130-A (PDiT 1.0) and http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3 (PDT 3.0).

[10] see http://ufal.mff.cuni.cz/pdit and http://ufal.mff.cuni.cz/pdt3.0/

phenomenon	**PDiT 1.0**	**PDT 3.0**
DISCOURSE RELATIONS		
explicit discourse relations	yes	yes – updated
- explicit connectives	yes	yes – updated
- semantic types	yes	yes – updated
list structures	yes	yes – updated
headings	yes	yes – updated
captions	yes	yes – updated
metatexts	yes	yes – updated
alternative lexicalizations	no	preliminary
genres	no	yes
COREFERENCE and BRIDGING		
grammatical coreference	yes	yes
textual coreference (extended)	yes	yes – updated
coreference of pronouns of 1st and 2nd person	no	yes
bridging relations	yes	yes – updated

Table 2.1: Discourse-related phenomena annotated in the PDiT 1.0 and in the PDT 3.0

2.3 Related Research

In this section, we first describe important work in the area of text studies in the Czech linguistics. Subsequently, we introduce influential approaches to discourse structure and coherence in international linguistics. Given the rich number and diversity of the international approaches, frameworks and studies, we have to be very selective and focus on those most relevant for this thesis. They are the Rhetorical Structure Theory (RST) and the Penn Discourse Treebank (PDTB) lexical approach. We focus on these ways of discourse analysis because they substantially influenced the development of views on analyzing discourse in general, and because they were implemented by corpus methods on authentic texts.

Other influential studies and frameworks regarding discourse analysis were not entirely omitted: some were mentioned earlier in Section 2.1 and some appear further on relevant places throughout the thesis.

2.3.1 Discourse in Czech Linguistics

In the Czech linguistic tradition, the first studies relating to some aspect of discourse appear long before the establishment of text linguistics as a separate discipline. In principle, it is research concerned with the structuring of given and new information, first within a sentence, later also in discourses. Of international importance is the work of V. Mathesius and the Prague Linguistic Circle, mainly the contrastive studies on word order and the information structure of a sentence, which is determined by both syntactic properties of the language and by contextual relations (Mathesius 1939, 1943). The early findings were later extended in Brno by J. Firbas and his notion of functional sentence perspective (Firbas 1974) and in Prague by P. Sgall and E. Hajičová's studies on topic-focus articulation (Sgall et al. 1973, Sgall et al. 1986) and the related notion of text topics and salience (Hajičová 1993). F. Daneš explores thematic progressions (*tematické posloupnosti*) in a text (1968) and analyzes in-depth phenomena on the border of syntax and discourse structure (1985).

A different viewpoint on discourse phenomena is represented by the development of stylistics, one branch of which also has roots in the Prague structuralist school; it emphasizes the functions of communication. K. Hausenblas (1964, 1971) proposes a detailed classification of discourses, their ways of construction and communicative functions.

A systematic account of the so-called "hypersyntax", along with a first complex proposal of discourse-related terminology, is offered by J. Hrbáček (1994).

There are, of course, a number of other Czech linguists contributing to discourse topics from various perspectives, J. Kořenský (1992), J. Tárnyiková (2002) and P. Šaldová (2002) to name but a few. Yet, longtime attention has been paid to one specific area: to discourse connectives and other discourse markers. These expressions have been in the focus of attention in the studies of J. Hoffmannová (mainly 1983 and 1984) where the ability of the connective expressions to fulfill different functions and their relation to unexpressed contents and pragmatic uses are highlighted. Further contributions, mainly to the definition of the category or individual case studies are offered by F. Daneš (1985), P. Adamec (1995), I. Kolářová (1998, 2002), Z. Bedřichová (2008), P. Jínová (2011); in Slovak linguistics especially by J. Mistrík (e.g. 1975). One of the most in-depth studies on discourse connectives is given in the monograph of O. Pešek (2011), which targets an argumentative subgroup in Czech and in French.

2.3.2 Discourse on the International Scene

In a nutshell, the leading approaches in discourse analysis, which focus on corpus and/or computational processing, access the issue from two main perspectives: the so-called “global” and “local” discourse structure modeling. In other words, the former approaches access discourse phenomena from the top, representing a whole document as a single connected structure (also referred to as a “deep” discourse parsing), and the latter ones access discourse phenomena from the syntactic perspective, looking for similar patterns in discourse (“shallow” discourse parsing). The project this thesis is concerned with, the Prague Discourse Treebank (PDiT 1.0), belongs to the latter perspective.

The most influential frameworks among the former are the Rhetorical Structure Theory (RST, Mann and Thompson 1988), the Discourse Graphbank (Wolf et al. 2005) and the Segmented Discourse Representation Theory (SDRT, Asher and Lascarides 2003).

The latter, “local” direction of discourse analysis is best represented by the lexically grounded approach of the Penn Discourse Treebank (for English, PDTB, Prasad et al. 2008), which accesses discourse relations in the first place by searching for their lexical anchors – discourse connectives. It also does not make any claims about the shape of the overall discourse structure.

In the following paragraphs, we will describe the RST and the PDTB ways of discourse analysis. We briefly characterize their starting points and describe their contributions to acquisition of linguistic knowledge from annotated data as well as to automatic processing. Also, the following introduction opens some general issues one encounters while developing such a theory. We discuss these issues in the last part of this section and we point out what lessons can (or must) be learned from the previous research.[11]

2.3.2.1 Rhetorical Structure Theory

The Rhetorical Structure Theory was originally developed by Bill Mann, Sandra Thompson and Christian Matthiessen at the University of Southern California in the eighties (Mann and Thompson 1988) with the intention to model text coherence in order to study computer-based text generation. The main principle of RST is the assumption that coherent texts consist of minimal units, which are linked to each other, recursively, through rhetorical relations and that coher-

[11] We do not pay special attention to the Segmented Discourse Theory in this section, even though it certainly deserves to be mentioned. We have already offered a detailed description of this theory and also an analysis of a Czech text within this framework in our diploma thesis (Mladová 2008a, pp. 33–40).

ent texts do not exhibit gaps or non-sequiturs (Taboada and Mann 2006). The RST, as a linguistic theory nonetheless independent of its computational uses, represents the whole text document as a single interconnected structure. Basic features of these structures are the *rhetorical relations*[12] between two *textual units* (smaller or larger blocks that are in the vast majority of cases *adjacent*) and the notion of *nuclearity*. For the classification of RST rhetorical relations, a set of labels was developed, which originally contained 24 relations, but the authors themselves add that it is an open set "susceptible to extension and modification for the purposes of particular genres and cultural styles" (Mann and Thompson 1988, p. 250). The type of a rhetorical relation is defined with respect to the author's intended effect on the reader together with the application of the principles of nuclearity (cf. Footnote 14). The *nuclearity* in RST roughly corresponds to the subordinate and coordinate syntactic relations in a language (Matthiessen and Thompson 1988). *Nucleus* is the one of the two connected text spans that represents more essential information for the text's purpose. (That means, from the syntactic viewpoint it is in principle the main, governing clause.) The other text unit, which brings rather background or supplementary information, is called the *satellite*. So, for instance, if a claim is followed by an evidence for the claim in the RST analysis, the relation will be labeled *Evidence* with the claim as a nucleus and the evidence for the claim as a satellite, cf. Example (4) from an online RST analysis.[13]

(4) *Darwin is a geologist.* [claim = nucleus] *His work contributed significantly to the field.* [evidence = satellite]

However, multi-nuclear relations also can be found, e. g. *Contrast* or temporal *Sequence*, cf. Example (5) from the same source.

(5) *One agent pointed to a massive chandelier* [nucleus] *and asked, "What would you call that in England?* [nucleus]

Both for nucleus and satellite, there can be pragmatic constraints on their realization that help define the relation holding between them.[14] Another important component of RST is the hierarchical organization of text units: rhetorical relations may enter recursively into new relations.

[12] also referred to as coherence relations or discourse relations in other theories

[13] http://www.sfu.ca/rst/pdfs/rst-analyses-all.pdf

[14] An example, again on the relation *Evidence*: the constraint on nucleus (N) is that the reader might not believe the N to a degree satisfactory to the writer. The constraint on satellite (S) is that the reader believes S or finds it credible. Hence, the constraint on both N and S, and so the definition of the relation itself, is that reader's comprehending S increases reader's belief of N.

All these features, the text units, the rhetorical relations between two adjacent units, the nuclearity principle and the recursion make it possible to represent a text document as a single tree-like structure.

The RST itself, as one of the first thorough attempts in modeling coherence relations, has gained great attention. Since its very beginning, there have been lots of reactions, it was further developed and tested, language corpora were built with RST-like discourse annotation (for instance, for English on a portion of an American business weekly Wall Street Journal – under the name RST Treebank (Carlson et al. 2002); for German the Potsdam Commentary Corpus (Stede 2004, Stede and Neumann 2014)). Also, in some of its theoretical claims, it was repeatedly opposed. The authors themselves decided to sum up the discussions twenty years later in two overview articles, one about the theory itself (Taboada and Mann 2006a), second about the applications inspired by the RST (Taboada and Mann 2006b).

2.3.2.2 The Penn Discourse Treebank

Since 1998, B. Webber and A. Joshi and their research team at University of Pennsylvania have been developing a lexically based model of discourse. Their analysis of discourse relations consists primarily in finding and analyzing lexical cues of discourse coherence as “anchors” of discourse relations. Such a cue, a *discourse connective*, is defined as a discourse-level predicate opening positions for two discourse arguments (two propositions, events, situations), cf. Webber et al. (1999). In the Penn Discourse Treebank annotation scheme, discourse connectives include coordinating conjunctions (apart from those coordinating mere sentence participants like “mum and dad”), subordinating conjunctions and discourse adverbials. A given set of approx. 100 types of discourse connectives was then manually annotated on the English texts of the business daily Wall Street Journal (henceforth WSJ).

Apart from connectives, the two discourse arguments of a discourse relation (and their extent) and the semantic type (sense) of a discourse annotation were annotated. Discourse arguments in the Penn Discourse Treebank are outlined as linguistic realizations of abstract objects (Asher 1993), prototypically predications with finite verbs, but also gerunds and nominalizations. As a convention, the argument containing the connective is marked as Argument 2, the other as Argument 1, disregarding its location.

For ascribing semantic categories to single discourse connectives in a context, a set of 30 semantic labels was developed, organized in a three-level hierarchy, cf. Figure 2.2 (Prasad et al. 2007). On the most general level, the class level, there are four semantic categories: TEMPORAL, CONTINGENCY, COMPARISON

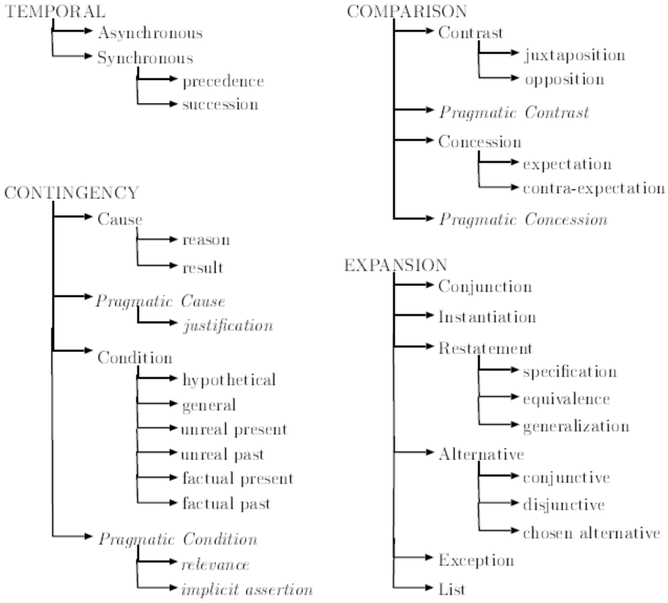


Figure 2.2: The PDTB sense label hierarchy

and EXPANSION. On the second level, the type level, there are further 16 categories (types) and on the third, most fine-grained level, some of the types are further sub-classified into subtypes.
In 2004, the first version of Penn Discourse Treebank was released (Miltsakaki et al. 2004). The second release of the PDTB four years later includes annotation of the ca. 49,000 sentences of the WSJ part of the Penn Treebank (PDTB 2.0, Prasad et al. 2008). Apart from explicit connectives, other phenomena have been annotated in this version, mainly implicit relations and connectives and attribution.

(i) Implicit connectives: discourse relations that are not realized by explicit DCs must be inferred by the reader. “In the PDTB, such inferred relations are annotated by inserting a connective expression called an “Implicit” connective that best expresses the inferred relation.” (Prasad et al. 2008). The implicit connectives were inserted into slots between two adjacent sentences, with the exception of paragraph boundaries. Where no appropriate implicit connective could be provided, the annotators could use three distinct labels (Prasad et al. 2008, p. 2963):

“AltLex” (alternative lexicalization of a connective) was used for cases where the insertion of an implicit connective would lead to a redundancy since the relation is signaled by some non-connective expression (cf. Example (6) from the PDTB).

(6) *Ms. Bartlett's previous work, which earned her an international reputation in the non-horticultural art world, often took gardens as its nominal subject.* AltLex: **Mayhap this metaphorical connection made the BPC Fine Arts Committee think she had a literal green thumb**.

"EntRel" (entity-based relation): was used for cases where only an entity based coherence relation could be perceived between the sentences (cf. Example (7) from the PDTB):

(7) *Hale Milgrim, 41 years old, senior vice president, marketing at Elecktra Entertainment Inc., was named president of Capitol Records Inc., a unit of this entertainment concern.* EntRel: **Mr. Milgrim succeeds David Berman, who resigned last month.**

Finally, "NoRel" (no relation) was used for cases where no discourse relation or entity-based relation could be perceived between the sentences (cf. Example (8) from the PDTB):

(8) *Jacobs is an international engineering and construction concern.* NoRel **Total capital investment at the site could be as much as $400 million, according to Intel**."

(ii) Attribution, in the PDTB terms, is "attributing beliefs and assertions expressed in text to the agent(s) holding or making them" (Prasad et al. 2007, p. 40), cf. Example (9) from the PDTB. In this example, the attribution clause is highlighted in bold, whereas the attributed content, in this case a direct speech, is highlighted in italics. A closer description of the annotation of attribution follows in Section 4.2.1.

(9) *"When the airline information came through, it cracked every model we had for the marketplace,"* **said a managing director at one of the largest program-trading firms.**

The PDTB-style connective/argument analysis has become very popular, also because such an analysis requires less interpretation and pragmatic inference than the RST analysis. The PDTB authors also claim that their approach is theory-neutral, independent from any syntactic theory, and as such can be transferred to other languages.

2.3.3 Recent Discourse Annotation Projects

In this section, we give an overview of annotation projects portraying discourse relations. First, we list corpora for English (arisen from different perspectives), and then corpora for languages different than English. The latter are mostly projects finished in recent years or running projects, and they nicely illustrate how the general interest in discourse-annotated language resources increases in the field of corpus and computational linguistics:

Discourse-annotated corpora for English:

- The RST-Treebank (Carlson et al. 2002)
- Discourse Graphbank (Wolf et al. 2005)
- The Penn Discourse Treebank 2.0 (Prasad et al. 2008)
- The BioDiscourse Relation Bank (BioDRB, Prasad et al. 2011)
- RST Signalling Corpus (Das et al. 2015)

Discourse-annotated corpora for languages other than English. Most of them follow in some way or another the annotation principles introduced by the PDTB team (Prasad et al. 2008):

- Hindi Discourse Relation Bank (HDRB, Kolachina et al. 2012, Oza et al. 2009).
- The Leeds Arabic Discourse Treebank (Al-Saif and Markert 2010)
- PDTB-style annotation of Chinese (Zhou and Xue 2012)
- Turkish Discourse Bank (Zeyrek et al. 2010)
- LUNA: PDTB-style annotation of Italian spoken dialogs (Tonelli et al. 2010)
- Potsdam Commentary Corpus (Stede 2004, Stede and Neumann 2014) – German
- French Discourse Treebank (Danlos et al. 2012)
- Tüba-D/Z Treebank (Gastel et al. 2011, Versley and Gastel 2013) – German
- AnnoDis (Afantenos et al. 2012) – French
- RST Basque Treebank (Iruskieta et al. 2013)

2.4 Theoretical Issues in Discourse Processing

There are many theoretical issues in conception possibilities of discourse processing arisen from the existing projects and their trickier parts. Tracking the existing discussions offers useful ideas of what can be achieved when developing a theory for discourse analysis or a corpus tagged for discourse phenomena, and also, what are the possible risks. The most important is naturally the purpose of the intended analysis – a strictly formal account would differ from a corpus-driven study or from a scenario for an annotation project.

The basic question how to represent discourse structure can be decomposed to smaller sub-questions. They are the basic ones and they must be answered in any discourse theory: What is a discourse-level unit? How the discourse-level units are connected, i.e. what is the nature of the discourse (coherence) relations among these units? Can they be described with formal means or are they rather a subject of psychological judgments? What are their classification criteria and their number? In what way do they participate in establishing discourse coherence and to what extent do they interact with other means of coherence?

In the following paragraphs, we discuss some of these theoretical questions and bring to light some of the fundamental views from the literature.

One of the most burning points of discussion, which is worth addressing more thoroughly, is the question of adequacy/sufficiency of representation of a discourse structure as a tree graph, as used in the RST. Linguistically, the strong constraints of a tree (no crossing edges, one root, all the units interconnected etc.) gave rise to a search for counter-arguments and counter-examples in real-world texts.

One direction pointed out that not only adjacent text units exhibit coherence links and that there are even cue phrases on the surface, which connect non-adjacent units and thus support the claim that a tree graph is too restricted a structure for an adequate discourse representation (Wolf and Gibson 2005). Therefore, more complex graphs with crossings and overlaps should be adequate for modeling discourse structure. This argumentation resulted in the creation of the Discourse GraphBank (Wolf et al. 2005; Wolf and Gibson 2005), a resource of English WSJ texts annotated in a similar fashion like the RST, but with the main diverging principle of relaxing the “tree-ness” constraint on the resulting representations.

Webber et al. (2003) and Lee et al. (2006) stated that cue phrases connecting non-adjacent units can be treated as anaphoric and thus they are of different nature and they do not violate the basic notion of a tree structure.

Another widely discussed property of RST is the representation of the whole document as one hierarchical structure. This concept has a strong potential in the possibility to demonstrate the composition of smaller blocks of the text, as well as to get to the more general and more important text contents and relations between them all the way up in the tree (following the nuclearity principle). The strong constraints have moreover the advantage that they hold the model in one piece while still enabling an elaborate, well considered analysis. The question asked at this point is, whether such a level of description is not too abstract to be agreed on by the analysts and, when we want to use it computationally, implemented by a machine. The opponents of RST, but also the authors themselves mention the rising degree of ambiguity in the interpretation of larger texts. We could say that this is a general issue encountering any attempt in such a complex task as text analysis. But the fact that the original RST made a large step away from the linguistic form, working directly and in a single representation with semantics, text topics and intentions, is a possible disadvantage for a reliable description by corpus methods. This fact triggered interesting further work in this area – e.g. on the nature, basic properties and classification criteria of discourse relations. We will address these questions in Section 2.4.2 below.

2.4.1 Layers of Discourse Analysis

Before turning our attention to the properties of discourse relations, let us make a short digression and mention another conclusion that can be drawn from the RST analysis. It concerns the need of a multilayer (multidimensional) analysis of discourse. The completeness (one schema application contains the entire text) and the complexity of the tree representation of discourse in RST can be seen as piling up several types of linguistic information for the purpose of getting a single structure. The hypothesis that in a coherent text each unit must be somehow linked to the others, is widely accepted, but the RST analyses do not tell us explicitly that there are more ways in which a text holds together, converting any such diverse information into one of the 24 coherence relations. A different, multilayered approach to discourse structure was first proposed by Grosz and Sidner (1986), who also show that there are different structures of a text (linguistic structure, intentional structure, attentional state) that together contribute to text coherence. They are interlinked and influence each other, but nevertheless should be recognized and held apart in the analysis. Similarly, Stede (2008) offers a division of RST analysis into several levels according to the nature of the coherence relations (referential, thematic, conjunctive[15] and intentional structure). In

[15] "*Conjunctive relations* are links that can be read off the text surface without performing "deep" inferences; these relations can be directed but they do not assign different degrees of prominence to the relata. Crucially, in MLA [Multilayer Analysis, LP] it is also possible that adjacent text segments are not linked by any such relation." (Stede 2008, pp. 319–320)

accordance with these views, we, too, argue that a multilayer analysis of discourse is needed. It should enable us to model in what way, with what frequency and intensity the various discourse-related phenomena take part in creating coherence. Surely, during the flow of the text, the extent with which several language means participate in linking the elements can be very different. Thus, we are convinced that in any annotation scheme for discourse phenomena, apart from discourse relations – the research of which will nevertheless remain the main objective of this thesis – there should be a separate layer of coreference analysis, a layer of intentions, of the text topics and their salience, of bridging relations, an analysis of graphical segmentation of the written texts, for spoken texts a layer of prosody analysis, and maybe other.

2.4.2 The Nature of Discourse Relations

The basic characteristics of discourse relations given in Section 2.1.4 explains the general notion shared by the discourse-oriented linguistic community. However, it does not say enough about the nature and properties of these relations in order to represent discourse structure. In the following paragraphs we describe the most pressing theoretical questions about the nature of discourse relations. It is the "semantic" and "pragmatic" dichotomy (Section 2.4.2.1), the possibilities of a definition of discourse relations by formal means (Section 2.4.2.2), the relation of discourse relations to sentential syntax (Section 2.4.2.3), the notion of nuclearity and (a)symmetry of discourse relations (Section 2.4.2.4) and some thoughts on granularity of semantic classification of the relations for annotation (Section 2.4.2.5).

2.4.2.1 "Semantic" and "Pragmatic" Relations

One of the most discussed properties of discourse relations is their "semantic" or "pragmatic" nature, in other words, the question where the source of coherence comes from, or what is actually related – propositions, inferences, illocutions, etc. The commonly used distinction "semantic" vs. "pragmatic"[16] is a little confusing, as the relations are always semantic but they either hold between text contents or between materials inferred. These two types of relations are very much interconnected and yet very different. One possible way to capture this issue theoretically is aptly explained by Kehler (2002), who offers a cognitive viewpoint for the interpretation of discourse relations and for coherence theory

[16] This distinction is used also in the PDTB sense taxonomy. In the PDTB, four pragmatic senses are distinguished and annotated: pragmatic cause, pragmatic condition, pragmatic contrast and pragmatic concession.

in general. According to him, in order to understand any real-world situation or discourse we perform a number of inferential processes. The degree of coherence depends on the amount of material inferred. If we are unable to infer any adequate piece of information (from the context or from our general world knowledge), the discourse/situation seems incoherent. We hereby satisfy the “desire to coherence” (Kehler 2002, p. 14), the need to resolve coherence. Kehler calls this process “coherence establishment”.

As for the “semantic”/ “pragmatic” distinction, Kehler agrees with the definition of Sanders et al. (1992): “A relation has a semantic source of coherence if the segments are related at the level of propositional content, whereas the source of coherence is pragmatic if they are related at the level of illocutionary meaning.” (Kehler 2002, p. 27). This distinction is also connected to inferential processes needed to establish a coherent relation: To be able to interpret a pragmatic meaning, one has to infer the right illocutions or unexpressed contents. According to Kehler, the “semantic”/“pragmatic” distinction is often less clear than other coherence features. In fact, there is also a three-way division of the “pragmatic” relations to *content*, *epistemic* and *speech act* readings (Sweetser 1991), demonstrated by the three following examples according to Sanders (1997), respectively. In the epistemic reading (11), writer’s reasoning is involved in the relation, (i. e. writer’s conclusion, that *John loved Mary* from the premise that *John came back*) and in the speech act reading (12) the causality holds between a speech act and the speaker’s justification of performing it. This sub-classification within the pragmatic domain to epistemic and speech act is used by some of the newest annotation projects, e. g. for Italian dialogs (Tonelli et al. 2010) – spoken dialogs appeared to offer a higher number of pragmatic usages of connectives than the written texts, or in the Hindi Discourse Relation Bank (HDRB, Kolachina et al. 2012; Oza et al. 2009).

(10) *John came back because he loved her.*

(11) *John loved her, because he came back.*

(12) *What are you doing tonight, because there’s a good movie on.*

There are, of course, many other attempts to further categorize the pragmatic domain, their adequacy being more or less supported by corpus data. Next, we will mention two such studies that might be useful to follow in future:

The authors of HDRB describe a situation where a whole proposition must be inferred to establish the coherence relation; they call it *pragmatic relation* *at propositional level* (d)[17]. The examples from Hindi data appear somewhat difficult, but a clear example can be found in Czech (taken from Hrbáček 1994):

(13) *Na vysokou školu se nehlásil. Stejně by se nedostal.*

He did not apply for the university. He would not pass the entrance exams, anyway.

Here, the whole proposition in the sense of *I kdyby se hlásil* (*Even if he applied*) represents the Argument 1 of the discourse relation. This third pragmatic subtype was included in the HDRB annotation scenario.[18]

Another interesting study of semantic and pragmatic nature of discourse relations was carried out by Robaldo et al. (2010), although they do not explicitly use the dichotomy "semantic"/"pragmatic". Their analysis of connectives with concessive meaning from the PDTB brought insight into different sources of expectations, the denial of which creates a concession. They found four different sources of expectations, i. e. different types of inference processes. Among them, only one can be seen as (denied) semantic Causality (Example (14)). Other three types, Implication (15), Correlation (16) and Implicature (17) require more complicated inferential processes (or, we can say, require to infer more material) to establish a coherent discourse and so they can be treated as pragmatic. In fact, Implicature does not convey a concessive meaning at all.

(14) *Although Greta Garbo was considered the yardstick of beauty, she never married.*

(15) *Although he does not have a car, he has a bike.*

(16) *John will finish his report, but he'll do it at home.*

(17) *Although it is not the first company to produce the thinner drives, it is the first with an 80-megabyte drive.*

The last example brings us to a different problem, which is nevertheless worth addressing in this section. Let's demonstrate it on the following three (invented) examples with a typically conditional *if* as a discourse connective. Under (18),

[17] "The propositional subtype involves the inference of a complete proposition. The relation is then taken to hold between this inferred proposition and the propositional content of one of the arguments". (Oza et al. 2009)

[18] However, when working with real-text data, the situation is sometimes so complex that the annotator cannot really say what syntactic structure the inference they make actually forms. From the nature of discourse units could be concluded that what is inferred is always a whole proposition. The subclassification of the pragmatic domain is in many discourse projects still a matter to be further investigated.

a trouble-free semantic relation of condition holds between the two clause contents. Under (19) some shifting of the conditional meaning happened that is needed to be understood to establish a semantic relation. It is a type of inference in sense of "*for the case* that you need some help, *you should know* that I will be next door". According to our approach, this case of indirect condition (Quirk et al. 2004, pp. 1088–1089) is already treated as a pragmatic condition – the *if*-connective does not connect directly the propositions represented by the two clauses. Rather, it signals the condition under which the speaker makes an utterance. Still, both formally and semantically, there is a conditional meaning and we can find many more examples of semantic relations of condition between inferred materials of any type. In the third example (20), however, the *if*-connective connects two contents that can be under any usual circumstances hardly interpreted as a condition. The rather untypical usage of *if* is nevertheless quite typical in these particular constructions of confrontation (with a slight tinge of gradation), easily replaceable with *whereas* or *while* (such structures cf. Quirk et al. 2004, p. 1087).

(18) *If you exercise a lot, you will win the contest.* LP

(19) *If you need some help, I will be next door.* LP

(20) *Jestliže včera Sparta hrála špatně, dnes to byla katastrofa.* LP

If Sparta played poorly yesterday, today it was a disaster.

In (20), a typically conditional connective (formal perspective) signals a contrastive meaning (semantic perspective). The point is that cases such as (20) should not be treated as pragmatic conditions. The formally conditional (and also syntactically subordinate) relation can mislead to some pragmatic interpretation of the condition. That is wrong, there is obviously no condition involved in such cases, neither between the very contents of the clauses, nor between any possible inferences or illocutions.[19]

To sum up, we have addressed the issue of semantic and pragmatic discourse relations to specify where the source of coherence comes from and to avoid confusion which may arise from the terminology. In some approaches, the so-called "false" or pragmatic (intra-sentential, mainly dependent) relations are named according to the formal perspective, according to the prevailing meaning of the

[19] Subordinate conjunctions in constructions with non-typical meanings, as demonstrated in Example (20), were targeted in our joint study with P. Jínová and J. Mírovský (Jínová et al. 2013).

connective.[20] In this view, Example (17) would be called a *false concession* and (20) a *false condition.* In our semantically based approach to discourse analysis, in contrast to the form-based terminology, we treat these cases as (semantic) contrasts/confrontations disregarding the prevailing meaning of the connective.

2.4.2.2 Are Discourse Relations Formally Definable?

Although it goes beyond the scope of this thesis to elaborate on this issue in depth, we find important to point out what appears to be the essential problem behind the formal accounts of description of discourse relations. Let us get there via the already mentioned cognitive approach by A. Kehler, as his classification of discourse relations demonstrates where the difficulties of formal accounts start.

According to Kehler (2002, p. 15, p. 26) we only perform a certain small number of cognitive processes in order to identify a discourse relation in our mind; these are the same processes that are also familiar operations from artificial intelligence. They determine Kehler's taxonomy of discourse relations, which is inspired by David Hume (1748) and uses three major categories: For the *Resemblance* category, the constraints are defined with help of set membership and relations among the subsets in sense of properties of individuals and sets involved – contrasting, comparing, exemplifying, drawing parallels etc. This should be a demonstration of our general cognitive ability to reason analogically (Kehler 2002, p. 18). The *Cause-Effect* category is based on implication, not in the strictly logical sense, but rather translated as "B could plausibly follow from A". The third and last category of *Contiguity* "is a bit murkier" (p. 22), as it expresses a sequence of eventualities centered around some system of entities, and so it requires to employ knowledge gained from human experience. So Example (21) from Samet and Schank (1984) is perfectly coherent despite the amount of material not mentioned at all:

(21) *Larry went into a restaurant. The baked salmon sounded good and he ordered it.*

Kehler comments on the *Contiguity* category as definable "in less formal terms than the others because precise constraints that utilize this knowledge prove difficult to state explicitly" (2002, p. 22). We would need to successfully model a whole semantic representation of event-types that typically happen, in a certain order, and are expectable in real world. The encoding of such knowledge seems to be the core of the problem of any formal treatment of discourse structure,

[20] In Czech syntactic description the phenomenon is prototypically called *nepravé věty vedlejší* (*false dependent clauses*). Also, in the tectogrammatical annotation of the PDT, they were annotated according to their formal structure rather than according to their semantics.

disregarding whether theoretical mathematical definitions or NLP applications are concerned. Other discourse theoreticians came to similar conclusions, namely, that certain group of discourse relations can be fairly satisfactorily described by appropriate formal means, whereas other group is less open to these means. Sanders et al. (1992) use binary features to distinguish discourse relations in their taxonomy, the first feature being the *basic operation* – the distinction between causal (P → Q) and other relations. If the relation is not causal, it is additive (P & Q). The problem of defining additive relations in such a way, though, is that any two propositions (that are both true for the speaker) can be put together and seen as coherent. Here, the constraints for a formal definition appear too weak.

2.4.2.3 Correspondence of Discourse Structure to Sentence Structure and Semantics

One theoretical issue about the nature of discourse relations is of particular interest for this thesis – the (partial) correspondence of discourse structure and semantics to the structure and semantics of (within) a sentence. We mention it briefly here for the sake of completeness; in the practical part of this thesis, Section 4.2 is devoted to this particular topic. The fact that the discourse project in Prague is based on the previous annotation of underlying syntax reflects the basic assumption that, roughly speaking, the semantics within a sentence is the same as the semantics of discourse relations (cf. Section 3.1). Thus, for instance, a causal relation between a predicate verb and its dependent clause remains the same causal relation on the level of discourse analysis. Moreover, any causal relation between separate sentences expresses “the same” causality (cf. Jínová et al. 2014). When analyzing a language starting from the smallest units, from the phonological and morphological level all the way up, as it is the case not only in the Prague school, we can state that discourse relations, or at least some of them, are syntactically motivated and syntax-bound (e.g. the conditional meaning, cf. again Jínová et al. 2014): we cross the sentence boundary and find the same semantic patterns. From the opposite point of view, when we start analyzing discourse composition, we will sooner or later arrive to discourse-relevant intra-sentential phenomena. Thus, there is no doubt that sentence syntax and semantics are of a great relevance for discourse analysis. However, there are, of course, such discourse composition principles one hardly finds within a sentence. To sum up, discourse relations are syntax-bound or syntax-independent with a different degree according to their semantic properties.[21]

[21] Jínová et al. (2014) refers to a joint study with Jínová and Mírovský, it is an extended version of a 2011 research paper from the Dependency Linguistics Conference. The study

2.4.2.4 (A)symmetry and Nuclearity of Discourse Relations

Only a small step away from the correlation between syntactic structure and discourse structure is the question of symmetry/asymmetry[22] of discourse relations or, in a different viewpoint, also the notion of nuclearity (again, nucleus is the one of the two arguments which is more central to the author's purposes). Previous sections were devoted to the nature of discourse relations; here we actually speak about the nature of discourse arguments entering a discourse relation.

In the PDTB approach, the third level of the sense hierarchy refines the senses on the second level (Prasad et al. 2008), cf. Figure 2.2 in Section 2.3.2.2, but it also determines the nature of the discourse arguments. For instance, for the *Asynchronous* TEMPORAL sense, the two arguments are always *precedence* and *succession*. We call this relation asymmetric since the arguments have different semantic properties. The same case holds for the PDTB *Cause* (and for causality in general): one argument is always the reason and the other is always the result. The properties of the arguments are clearly set and clearly recognizable, no matter in which order the arguments appear. Such an obvious argument classification, however, is not possible with other types of relations, which we call symmetric. Confronting, contrasting, temporal synchronicity, conjunction, disjunction and equivalence bring together arguments that have the same semantic properties – the only way the arguments of these relations differ is their order of appearance in a text.

If a discourse relation is realized between a governing and a dependent clause within a single sentence (intra-sententially), the relation seems to be always asymmetric. The syntactic government/subordination signals a different semantic nature of the two arguments, and, in the RST terms, also nuclearity. The governing clause is then the nucleus and the dependent is the satellite. According to RST, the satellites can be in principle omitted while the main information of the relation is preserved in the nucleus. Multinuclearity is a phenomenon of certain paratactic structures: "In RST, parataxis is reflected in multinuclear relations, those where no span seems more central than the other to the author's purposes." (Taboada and Mann 2006a).

The symmetry and asymmetry of discourse relations, and also the notion of nuclearity bring consequences for designing a semantic classification of the relations. We have to carefully distinguish between refining discourse semantic types to further subtypes on the one hand, and between characterizing the arguments

demonstrates the different degree of syntax-boundness for three discourse relations: *condition*, *opposition* and *specification*.

[22] Symmetry and asymmetry of discourse relations are not understood as mathematical notions.

of the relations on the other. Moreover, the awareness about different semantic properties of arguments of different relations is useful for setting rules for annotating these arguments.

2.4.2.5 Granularity of Semantic Types

In the discourse-oriented literature, there is a huge discussion on the number and classification criteria for discourse relations. There are many proposals of sets of discourse relations, the number of which varies from two relations (Grosz and Sidner 1986; Polanyi 1988) to large, fine-grained sets (cf. the comparative study of Hovy 1990). Many of these taxonomies are proposed hierarchically, which makes even the very detailed relation sets convertible to few more general discourse categories. The apparent question here is what the most reasonable option for a representation is. The answer, again, is dependent on the purpose of the analysis. For manual data annotation and subsequent machine learning tasks, one risks facing data sparsity and lower inter-annotator agreement, if the label set is too rich. On the other hand, a rather modest set of relations can lead to omissions of important types of information. Is then, for this purpose, the middle way the best? The RST, as already stated, originally used a taxonomy of 24 relations (Mann and Thompson 1988, p. 250), which was further refined to 78 relations in 16 classes for the purposes of RST-Treebank annotation (Carlson and Marcu 2001). Wolf and Gibson (2005) use a set of 10 relations[23] in four general classes for the annotation of their Discourse GraphBank. The Penn Discourse Treebank 2.0 has been annotated with a set of 30 relations in four general classes in three-level hierarchy, the number 30 being the most detailed level. In Prague discourse annotation, the definite number of discourse relations assigned to the texts is 22 in four major classes. It seems a common empirical experience of those who work with discourse-aimed corpora that for the purposes of data annotation the discourse relations set should contain around 20 to 30 relations. This granularity, in our opinion, enables one to reach reasonable inter-annotator agreement and at the same time not to use too general, less informative categories. Also, from a cognitive point of view, one can imply that, in general, in order to understand texts uniformly, and as meaningful, coherent and unambiguous wholes, the readers make distinctions between the relations on approximately such a level of semantic granularity, no less detailed and no more detailed.

[23] altogether 11, with the *Same* relation (a continuation of a discontinuous argument)

3

Discourse Annotation in the PDiT 1.0 and the PDT 3.0

In this chapter, we describe the process of creation of the Prague Discourse Treebank 1.0 (PDiT 1.0), i.e. the discourse annotation of the Czech texts from the Prague Dependency Treebank. As the work on the project spanned across more than four years and went through different phases, there have already been published some work-in-progress reports, evaluations and first corpus-based studies during this period. Also, there are extensive annotation guidelines in English in the form of a technical report (Poláková et al. 2012b)[1]. The chapter is divided into several sections in which first theoretical starting points (Section 3.1), then the practical annotation process (Sections 3.2 to 3.7) and, finally, the evaluation of the annotated data (Section 3.8) are described in detail. Some parts of the description of the manual annotations in the PDiT are to some extent similar to the annotation guidelines provided in the technical report. However, the guidelines were put together before and during our real touch with the data. The following chapter of the present thesis, in contrast, offers an updated and summarizing look back on the processed and released treebank.

Throughout this chapter, we refer mostly to the PDiT 1.0 version of the annotation, as it is the first resource with this type of annotation and the first one publicly released. Where needed, we describe the adjustments and changes in the more recent data release within the PDT 3.0.

[1] The Czech version of the annotation guidelines is so far unpublished and it is available upon request.

3.1 Basic theoretical decisions

The idea to create a discourse-annotated corpus for Czech emerged as a possibility of testing the lexical approach of connective identification in the PDTB on a syntactically more complex (and typologically different) language (Lee et al. 2006, Footnote 1). For the Prague group, this was quite a natural step to do, as we had at our disposal a large, multilayer-annotated resource for Czech (PDT 2.5), the tectogrammatical level of which offered already some information possibly relevant for discourse annotation in the sense of PDTB. This was our basic assumption and a starting point at the beginning of the project: **A syntactico-semantic analysis of a sentence contains (retrievable) information about relations in discourse.** Or, in other words: **Certain entities and relations in a syntactico-semantic analysis of a sentence have corresponding counterparts in an analysis of discourse.** What kind of information this is, how it is represented in the Prague Dependency Treebank and in which way and to what extent it can be adopted and employed in building a discourse-annotated corpus was the main topic of our diploma thesis (Mladová 2008a). Later, the application itself was discussed in research papers by Mírovský et al. (2012) and Jínová et al. (2012), of which the author of this thesis was a co-author. Two main theoretical decisions for discourse representation in Prague are the following:

- inspiration by the PDTB lexical approach and annotation scenario
- annotation on syntactic trees of the tectogrammatical layer

Another, a rather practical decision is connected to the latter point, to the decision to annotate discourse directly on top of syntactic trees:

- two-phase annotation (first manual, then computer-aided)

The following two sections discuss the motivations behind these decisions, respectively.

3.1.1 Inspiration by the PDTB Approach

The approach of the PDTB group was reflected in the build-up of annotation scheme of the PDiT 1.0 in two main points: The first point is the basic concept of connective identification, the identification of the two arguments of the connective and the assignment of a semantic label to the relation signaled by the connective. This is the primary method adopted. The second point of inspiration by the PDTB was the shape of the hierarchy of sense tags for discourse. Here, the subsequent PDTB-like projects had the advantage to use the empirical experience made by the PDTB creators. So, for each of the projects, including the PDiT, there are some adjustments to the sense hierarchy. Mostly, the original division into four major semantic classes is preserved but within these classes, the repertoire of the relations varies.

In the PDiT 1.0 and the PDT 3.0, the annotations of discourse relations are limited to the relations expressed by explicit DCs (present on the surface); other tags (for implicit connectives, AltLex, EntRel and NoRel, cf. Section 2.3.2.2) between adjacent sentences were not assigned. Alternative lexicalizations (AltLex) were annotated in a preliminary fashion, with no sense assignment so far. Their thorough analysis, though, is a work in progress (Rysová 2012a). Entity-based relations (EntRel) are, in our view, a matter of coreference and bridging relations. As such, these relations are annotated in the PDiT 1.0 within another subproject (cf. Nedoluzhko 2011 and also Poláková et al. 2013).

Another phenomenon not annotated in Prague treebanks so far in comparison with the PDTB is attribution. We believe that this information can be at least partially obtained from syntactic features of the syntactic layers of PDT, e.g. attributes for direct speech, parentheses, verbal valency etc. (cf. Poláková et al. 2013).

A comparison of discourse phenomena annotated in the PDTB 2.0 and the PDT 3.0 is given in Table 3.1. We find it more convenient to refer here to the latest data version released, which is the PDT 3.0. For comparison of annotations in the PDiT 1.0 and the PDT 3.0, cf. Table 2.1 above.

phenomenon	**PDTB 2.0**	**PDT 3.0**
explicit DCs	yes	yes
implicit DCs	yes	no
Altlex	yes	preliminary
EntRel	yes	within coreference annotation
NoRel	yes	no
attribution	yes	no
list structures	as a sense tag	as a special type of structure
headings	no	yes
genres	yes	yes

Table 3.1: A comparison of discourse annotations in PDTB 2.0 and PDT 3.0

3.1.2 Annotating on Top of Syntactic Trees

The main motivation for carrying out the annotation of discourse phenomena on syntactic (tectogrammatical) trees was to preserve the connection with and information from the analyses of previous levels. The aim was to mine the treebank for all the already once manually annotated information that can be relevant for representing discourse structure. This is quite a unique approach among the similarly aimed projects[2] and it brings many (both linguistic and technical) advantages, but also some disadvantages. The main benefits are the following:

- Almost all syntactic counterparts of intra-sentential discourse relations are in fact already annotated within the tectogrammatical layer, which makes the information easily automatically retrievable for a discourse annotation.
- Many inter-sentential DCs were also marked within the tectogrammatical analysis. They are assigned the functor PREC (reference to PREceding Context). The PREC-functor represented a strong guide for the annotators and it also played a significant role in final checking procedures.
- A substantial advantage of using the tectogrammatical representation was the ellipsis restoration annotated on this layer. In particular, resolution of structures with elided verbs helped determine the borders and the extent of discourse arguments.
- The possibility to directly confront the syntactic structure of a discourse argument helped sort out such phenomena as parentheses, reporting clauses, appositions, coordinations of mere noun phrases etc.
- Of great advantage was also the possibility for the annotator to search for and visualize more linguistic phenomena at once. Specifically, when annotating discourse relations, the possibility to display also the annotated coreference chains and bridging anaphora is of great help. Some coreferential relations are known to be a distinctive feature for recognizing a DC or a discourse relation.

Disadvantages accompanying the annotation of discourse relations directly on the syntactic trees are mostly of a technical nature.

[2] The PDTB annotation was carried out on raw texts and only afterwards it was mapped onto the constituent trees of the original Penn TreeBank. In this sense, the treebank users do have the different annotation layers at once at their disposal. The main difference lies in the fact that the annotators could not use and be influenced by any syntactic information.

- To learn to read a tectogrammatical tree, a structure with quite a rich annotation and a high level of abstraction, takes a while. Also, marking the discourse relation and all its attributes in the tree representation requires concentration and it is a time consuming task.
- In the tree-mode of the annotation tool, large arguments cannot be displayed as a whole at one time[3]. Some adjustments of the tool were made to make the work with trees as comfortable as possible. A textual window is always used simultaneously with the tree-window.
- One methodological disadvantage of annotating discourse relations directly on syntactic trees is a possible restrictive thinking of the annotators in terms of (sub)trees. We were aware of such a tendency – tree structures intuitively underline the respective sentence boundaries, and, besides, possible places where the argument boundaries mismatch with the (sub)tree boundaries could be overseen. Also, looking for DCs in a tree representation is not easy. That is why we asked the annotators to work first with a hard copy and raw texts only, to find DCs for a whole document at a time, think of the relations, and only then they started to work with the tree representation.

3.1.3 Two-phase Annotation

The annotation of discourse relations in the PDiT 1.0 consisted of two phases:

In the first phase, the treebank was thoroughly manually processed, the annotators focused on inter-sentential discourse relations (relations between sentences) signaled by explicit discourse connectives. Intra-sentential relations were only marked manually in cases where the tectogrammatical representation did not convey a certain type of discourse semantics (Jínová et al. 2012), according to the annotation guidelines set for the discourse (Poláková et. al 2012b).

The second, subsequent phase focused on the remaining, so far unmarked intra-sentential discourse relations. We performed an automatic extraction of relevant syntactic features, namely those corresponding to some relations of syntactic dependency or coordination within a sentence, along with their connectives and arguments. These were then automatically mapped onto the discourse annotation. A detailed description of the annotation procedure in both phases is given in Section 3.3.

[3] Depending on the computer screen used, one can typically display up to 6 trees in a reasonable (readable) size.

3.2 Data Format, Annotation Interface and Querying

3.2.1 Data Format

The primary format of Prague treebanks is called PML (Prague Markup Language)[4]. It is an abstract XML-based format designed for annotation of linguistic corpora, and especially treebanks. PML-formatted data can be browsed and edited in a tree editor TrEd (cf. Section 3.2.2) and processed automatically using btred, a command-line tool for applying Perl scripts to PML data.

3.2.2 TrEd and Data Representation

For most types of manual annotation of the Prague treebanks, the annotation interface TrEd is used. TrEd is a fully customizable and programmable graphical editor and viewer for tree-like structures[5] (Pajas and Štěpánek 2008). It can be easily adjusted to a desired purpose by extensions that are included into the system as modules. The TrEd extension implemented for the purposes of discourse annotation on top of syntactic trees (cf. Mírovský et al. 2010) offers several specific features for this type of annotation:

- the creation of a link between the arguments of a discourse relation;
- exact specification of the arguments of the relation;
- assignment of a connective to the relation (or vice versa);
- assignment of additional information to the relation (semantic type etc.).

In the following paragraphs, we describe the PDiT 1.0 and PDT 3.0 data representation in TrEd (summed up according to Mírovský et al. 2010). The data representation meets the requirements on an annotation tool for discourse mentioned above. We provide this rather technical description of the data format because it enables the treebank users to query the treebank in an effective way. Linguistic characterization of the annotated phenomena is given in Sections 3.4 to 3.7 below.

Links between arguments: The annotation of discourse relations in the PDiT is performed on top of the tectogrammatical (deep syntactic) trees. A discourse relation is represented as an oriented link between two tectogrammatical

[4] http://ufal.mff.cuni.cz/jazz/pml/

[5] http://ufal.mff.cuni.cz/tred/

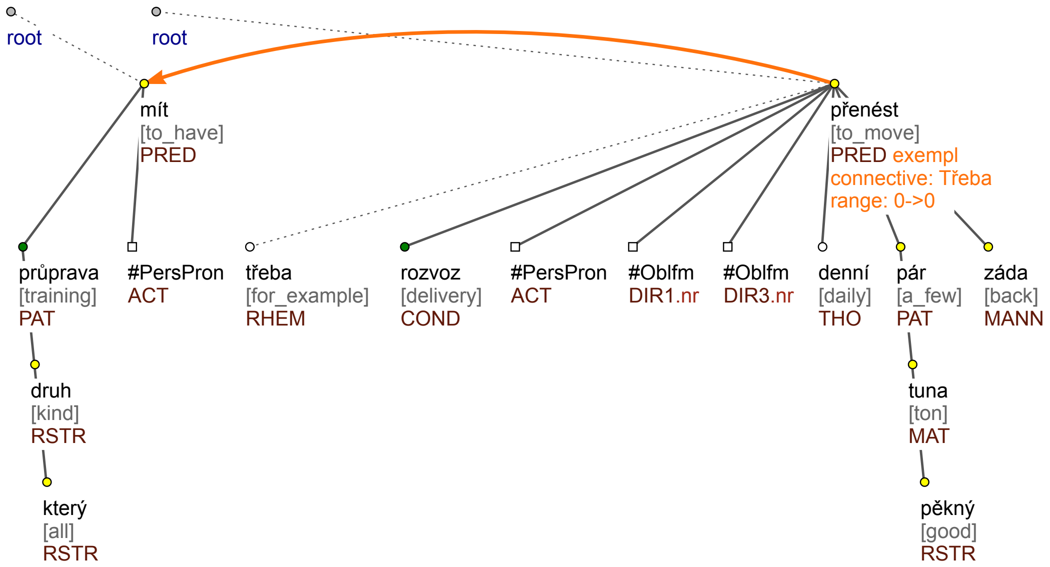


Figure 3.1: An arrow representing a discourse link

nodes in any trees (of a single document). The link is constituted by a dedicated attribute (*discourse/target_node.rf)* at the initial node of the relation, containing a unique identifier of the target node of the relation.[6] The link is depicted as an orange curved arrow between the nodes, cf. Figure 3.1. Although the arrow connects two nodes, it does not mean that the two nodes themselves equal the two arguments of the relation – cf. extent of the arguments in the following paragraph.

Additional information about the relation is also kept at the initial node of the relation – there is an attribute for the semantic type of the relation, an attribute for the source (annotator's initials – not included in the PDiT 1.0 release) and an attribute for annotator's comment (partially included in the PDiT 1.0 release).

Extent of the arguments: Usually, an argument of a discourse relation corresponds to a subtree of a tectogrammatical tree. As such it can be represented simply by the root node of the subtree. The convention is that the whole subtree is understood as a discourse argument. However, sometimes it is necessary to exclude a part of the subtree from the argument, sometimes the argument consists of more than one tree and sometimes it is even impossible to set the borders of the argument exactly. To allow for all these variants, each discourse link has two

[6] The data representation allows for several discourse links starting at a single node – there is a list of structured discourse elements representing the individual relations.

additional attributes specifying the range of the initial argument (the attribute *start_range*) and the range of the target argument (the attribute *target_range*). Both are stored at the initial node of the link. Their possible values are:

- "0" (zero) – the argument corresponds to the subtree of a given node;
- "n" (a positive integer) – the argument consists of the subtree of a given node and of *n* subsequent (whole) trees;
- "group" – the argument consists of an arbitrary set of nodes (details below); this option is used only if the previous options are not applicable;
- "forward" – the argument consists of the subtree of a given node and an unspecified number of subsequent trees; this option is used only if more specific options are not applicable;
- "backward" – analogically, the argument consists of the subtree of a given node and an unspecified number of preceding trees; this option is used only if more specific options are not applicable.

Groups: An argument of a discourse relation may consist of an arbitrary group of nodes, even from several trees. This fact is indicated in the range attribute of the relation (by the value "group"). Another attribute then tells which group it is. Groups of nodes inside one document are identified by numbers (positive integers). Each node may be a member of several groups; a list of identifiers of groups a node belongs to is kept at the node. Every group has a representative node – if a discourse link starts/ends at a group, it graphically starts/ends at the representative node of the group, which is the depthfirst node of the group belonging to the leftmost tree of the group. Figure 3.2 shows an example of a group annotation for the sentence in (22), the text segment belonging to the group is highlighted with a blue font.

(22) *K pěstování vědy je třeba nejen střecha nad hlavou, nějaké finance* (a **někdy jich je třeba dost**), ale především vědecký dorost.

For cultivation of science, it is necessary to have not only a roof over your head, some finances (and **sometimes there needs to be plenty**), but especially young researchers.

Connectives: A connective of a discourse relation is represented as a list of identifiers of (usually) tectogrammatical nodes that correspond to the surface tokens of the connective; the list is kept at the initial node of the relation. It often contains only one node, but sometimes it consists of several nodes. Some

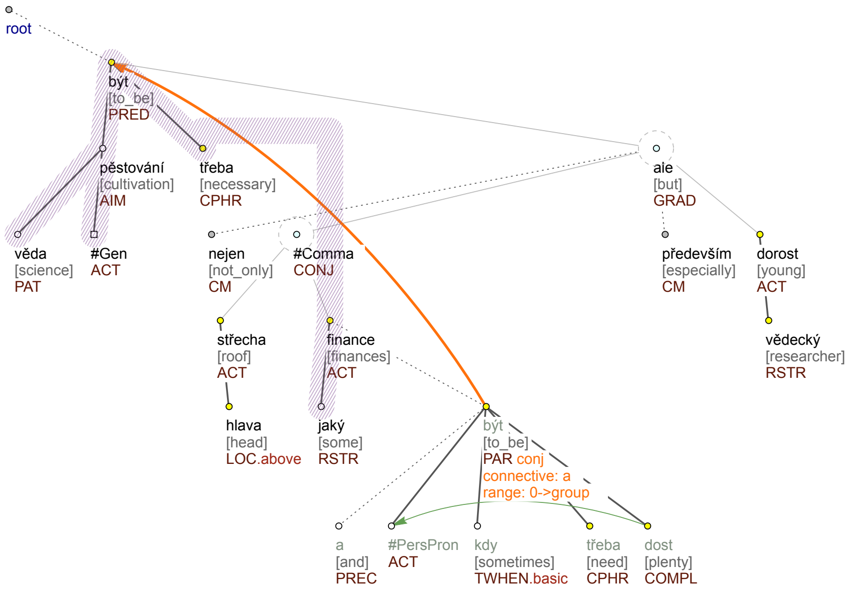


Figure 3.2: Example of a group annotation

connectives (e.g. punctuation marks) are not always represented on the tectogrammatical layer (at least not as a node). Therefore, identifiers of nodes from the analytical layer (surface syntax) are allowed as well.

List structures: List structures are enumerative constructions, annotated in the PDiT 1.0 and in the PDT 3.0 as independent compositional structures (for a linguistic description cf. Section 3.7.1). Their data representation is analogous to that of the discourse relations: by an oriented link from the root node of each list item to the root of the previous list item (for the first item in the list the target node is omitted). The attribute *discourse/type* of the link has the value "list" (while the value "discourse" indicates a discourse relation).

Other features: The TrEd tool incorporates also other features that make the annotation of discourse relations easier. Based on their preference, the annotators can annotate the relations either on the trees or on the linear form of the sentences in the text window of the tool. In the sentences, the tokens that represent the initial/target nodes of the relations are highlighted and easily visible. The annotators can also save space on the screen by contracting the trees, so that one node corresponds to one clause. Also, the attribute of annotator's comment (of an arrow or of a node) enables the annotators to comment on problematic cases.

In the PDiT 1.0, discourse-related annotation is captured mostly in a structured attribute *discourse* at the start node of the relation; additional annotation is captured in attributes *discourse_groups, discourse_comment* and *is_heading.*

In the PDT 3.0, there are small changes to the discourse attributes and their structuring:

A new attribute *discourse_special* is introduced, with possible values "heading" (for marking headings and titles of the corpus texts, it replaces the attribute *is_heading*), "caption" (for marking captions of photos, tables and charts) and "metatext" (for metatext information occurred by mistake during corpus compilation). The only document-level attribute is *genre* which captures the different types of genres of the treebank documents, newly annotated in the PDT 3.0. There are 20 possible values of the genre attribute (cf. Poláková et al. 2014).[7]

The overview of the attributes for both treebank versions follows:
Attributes applied both in the PDiT 1.0 and in the PDT 3.0:

- ***discourse/target_node.rf*** – id of the target node, or undefined if there is no target node (e.g. no hypertheme in a list structure)
- ***discourse/type*** – type of the arrow, two possible values: *discourse* (discourse relation), *list* (list entry)
- ***discourse/start_range*** – start range of a discourse arrow; for possible values cf. extent of the arguments above
- ***discourse/target_range*** – target range of a discourse arrow; for possible values cf. extent of the arguments above
- ***discourse/start_group_id*** – identifier of a group of nodes (positive integer) where the start_range of the arrow is set to "group"; individual nodes belonging to the group keep the group identifier in the attribute *discourse_groups*
- ***discourse/target_group_id*** – identifier of a group of nodes (positive integer) where the target_range of the arrow is set to "group"; individual nodes belonging to the group keep the group identifier in the attribute *discourse_groups*
- ***discourse/discourse_type*** – semantic type of a discourse relation, 23 values possible such as *equiv* (equivalence) or *conc* (concession)

[7] We do not target the genre annotation in the present thesis. A detailed description of the subject is given in Poláková et al. (2014).

- ***discourse/t-connectors.rf*** – list of ids of nodes from the tectogrammatical layer that represent a discourse connective
- ***discourse/a-connectors.rf*** – list of ids of nodes from the analytical layer that represent a discourse connective
- ***discourse/comment*** – annotator's comment of a discourse arrow (relation)
- ***discourse__groups*** – list of identifiers of groups the given node belongs to
- ***discourse__comment*** – annotator's comment of a node

Attributes applied in the PDiT 1.0 only:

- ***is__heading*** – set to ”1” at roots of subtrees representing article headings

Attributes applied in the PDT 3.0 only:

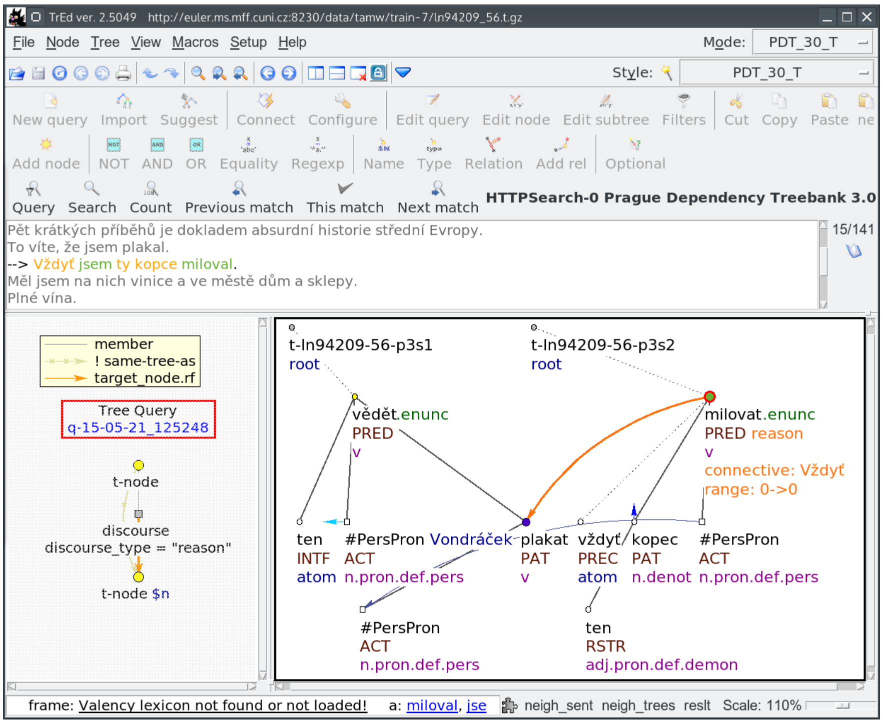
- ***discourse__special*** – marking of specific discourse phenomena, possible values “heading”, “caption” and “metatext”
- ***genre*** – genre type of a corpus document (document-level attribute).
  20 possible values, such as “news”, “essay” etc.

3.2.3 Querying: PML-TQ Search Engine

In the present thesis, for accessing any type of linguistic information annotated on the data of PDiT 1.0 and PDT 3.0, the search engine PML-Tree Query[8] was employed (Štěpánek and Pajas 2010). It is a powerful client-server based system designed specifically for querying all kinds of linguistically annotated treebanks (in the PML format). The server part performs the search and is implemented as a relational database. The client part provides a user interface and is implemented either as a TrEd extension or a web-browser based service. The TrEd client version has a more user-friendly interface and allows for a graphical creation of the queries but it requires an installation (of Perl, TrEd and the PML-TQ extension in TrEd). The web-based client[9] does not require any installation but it requires inputting a query in a textual form.

[8] http://ufal.mff.cuni.cz/pmltq/

[9] available from https://lindat.mff.cuni.cz/services/pmltq/



TrEd ver. 2.5049 http://euler.ms.mff.cuni.cz:8230/data/tamw/train-7/ln94209_56.t.gz
File Node Tree View Macros Setup Help
Mode: PDT_30_T
Style: PDT_30_T
New query Import Suggest Connect Configure Edit query Edit node Edit subtree Filters Cut Copy Paste ne
Add node NOT AND OR Equality Regexp Name Type Relation Add rel Optional
Query Search Count Previous match This match Next match
HTTPSearch-0 Prague Dependency Treebank 3.0
15/141
member
! same-tree-as
target_node.rf
Tree Query
q-15-05-21_125248
t-node
discourse
discourse_type = "reason"
t-node $n
t-ln94209-56-p3s1
root
t-ln94209-56-p3s2
root
vědět.enunc
PRED
v
milovat.enunc
PRED reason
v
connective: Vždyť
range: 0->0
ten
INTF
atom
#PersPron Vondráček
ACT
n.pron.def.pers
plakat
PAT
v
vždyť
PREC
atom
kopec
PAT
n.denot
#PersPron
ACT
n.pron.def.pers
#PersPron
ACT
n.pron.def.pers
ten
RSTR
adj.pron.def.demon
frame: Valency lexicon not found or not loaded!
a: miloval, jse
neigh_sent neigh_trees reslt Scale: 110%

Pět krátkých příběhů je dokladem absurdní historie střední Evropy.
To víte, že jsem plakal.
--> Vždyť jsem ty kopce miloval.
Měl jsem na nich vinice a ve městě dům a sklepy.
Plné vína.

Figure 3.3: The PML-Tree Query search engine

Figure (3.3) shows a snapshot of the PML-TQ extension in TrEd. The query represented graphically on the left bottom side defines two t-nodes (tectogrammatical nodes) connected with a discourse relation of the type “reason-result”, as stated by the value “reason” of the attribute *discourse_type* at the technical middle node representing the properties of the arrow. Another requirement set by the query is that the two connected nodes are not from the same tree, i.e. we are searching for an inter-sentential discourse relation of the type “reason-result”.

In the bottom right part of the figure, one of the results of the corpus search is displayed. The sentences represented by the two trees, along with the context, are depicted in the middle part of the figure. The two displayed trees represent Czech sentences that can be translated as: *Of course I cried. After all, I loved the hills.*

3.3 Annotation Procedure

As mentioned earlier, the annotation of discourse relations has proceeded in two phases. The first phase was manual and focused predominantly on inter-sentential relations; the second phase included automatic extraction of relevant syntactic features – thus focusing on intra-sentential relations. Both types of annotation underwent consistent checking procedures.

3.3.1 Manual Part

During the manual annotation phase, the annotators first worked with plain texts where they identified all instances of discourse connectives. This is a different approach from the one of the PDTB group, where an annotator went through all the occurrences of one connective type in the whole treebank, i.e. the annotator annotated for example "all the *becauses*". In such a way, the set of possible DCs is determined in advance – there is a list of expressions to be annotated. The Prague annotators had more responsibility in this respect, as they had to decide themselves if any expression in a given context functions as a DC, according to the criteria for DCs set in advance in the annotation guidelines. In this way, a discussion could arise whether a certain expression in a certain context actually fulfills the DC criteria. This approach may be less consistent as for the delimitation of the DC category but it provides some interesting linguistic material on the periphery of the category and makes its further research possible.

Only after having searched for DCs in their hard copies of the corpus texts, the annotators worked with the tree structures in TrEd. Having identified the connective, its two arguments (i.e. their extent) were set (creation of the discourse arrow), and to each such relation, one of the labels for semantic types was assigned.

Another difference in the process of annotation in Prague in contrast to PDTB poses the assignment of semantic labels (sense tags) to the relations. The sense tags in Penn Discourse Treebank 2.0 are organized in a three-level hierarchy with four top semantic classes, 16 types on the second and further 23 subtypes on the third hierarchy level (Prasad et al. 2008). The PDTB annotators were not forced to make the finest distinction (on the third, subtype level). A relation could also be annotated with two senses, forming a composite sense with a label combination from wherever in the hierarchy, resulting in 129 theoretically possible distinct sense tags. For this reason, some of the sense labels are very scarcely used, although they may be important for fine-grained distinctions in English. As confirmed by Meyer and Poláková (2013), this granularity level might not be useful for NLP uses of the data.

In the Prague semantic label system, there are 22 relation types in four general classes; the annotators had to choose one of the 22 types. In Section 2.4.2.5 above, we argue that such level of semantic granularity seems to be the best solution to avoid data sparsity on one hand, and not to lose relevant semantic information on the other.

Intra-sentential discourse relations, i.e. those that correspond to some syntactic relations already captured within the tectogrammatical analysis, were newly manually annotated only if their discourse semantics differed from the tectogrammatical interpretation. This is the case for pragmatic interpretations, finer subcategorization of adversatives etc. (cf. Jínová et al. (2012) and Section 3.6.5.3 on contrastive relations).

3.3.2 Computer-aided Part

The second, computer-aided part of PDiT annotation was based on extracting discourse-relevant information (presence of the relation, scope of the arguments, the connective(s), a semantic label) from the tectogrammatical layer of the PDT. The whole procedure, including some manual preparatory work mainly concerning temporal relations, is described in Jínová et al. (2012). As mentioned earlier, the tectogrammatical tree structures offer some types of information that can be transferred to the discourse-level annotation. In general, this concerns subordinate syntactic relations between clauses with labels like causality, conditionality, temporality, concession etc.; and coordinate syntactic relations between clauses within one sentence with selected coordinative labels like conjunction, disjunction, adversative meaning, confrontation etc. These relations were semiautomatically transferred to the discourse annotation (under the names vertical (subordinate) and horizontal (coordinate)). In a nutshell, all syntactic relations with a specific functor that were not annotated previously in the manual phase were transferred as follows (cf. also Table 3.2 with functor-to-discourse-type conversion):

If a tectogrammatical node represented:

- a finite verb with one of the temporal functors (TFHL, THL, THO, TSIN, TTILL, TWHEN), the node was annotated using the information from a manually pre-processed table (Jínová et al. 2012).
- a finite verb with one of the functors CAUS (cause), COND (condition), CNCS (concession), AIM (aim), CONTRD (contradiction) or SUBS (substitution), the node became a candidate for an automatically detected vertical discourse relation.

- a coordination node with one of the functors REAS (reason), CSQ (consequence), ADVS (adversative), CONFR (confrontation), GRAD (gradation), CONJ (conjunction) or DISJ (disjunctions), which coordinates (directly or transitively) finite verbs or other nodes with the functor PRED (predicate), the given node became a candidate for a horizontal relation.

The candidates for vertical and horizontal relations were checked for the presence of a previously manually annotated relation; if there was none, an automatic discourse relation was created, in a basic case of a vertical relation directly between the dependent and governing verbal nodes, and, in a basic case of a horizontal relation, between the members of the coordination. The treatment of more complex structures is also described in Jínová et al. (2012). Unlike tectogrammatical relations, discourse semantic relations in our approach do not reflect syntactic hypotaxis and parataxis (for details cf. Section 3.6.1). This is best demonstrated in the class of contrastive relations. For instance, as Table 3.2 indicates, the discourse type of *confrontation* may be represented by two different functors: CONTRD (contradiction) – syntactic subordination, and CONFR (confrontation) – syntactic coordination. In all cases, the connectives were detected automatically on the basis of other tectogrammatical and analytical (surface syntactic) attributes.

	Functor	**Long name**	**Discourse type**	**Long name**
vertical (dependency functors)	AIM	purpose	purp	purpose
	CAUS	cause	reason	reason-result
	CNCS	concession	conc	concession
	COND	condition	cond	condition
	CONTRD	contradiction	confr	confrontation
	SUBS	substitution	corr	correction
horizontal (coordination functors)	ADVS	adversative	opp	opposition
	CONFR	confrontation	confr	confrontation
	CONJ	conjunction	conj	conjunction
	CSQ	consequence	reason	reason-result
	DISJ	disjunction	disjalt	disjunctive alternative
	GRAD	gradation	grad	gradation
	REAS	causal relation	reason	reason-result

Table 3.2: Functor-to-discourse-type automatic translation table

3.4 Connectives in the PDiT 1.0

Discourse connectives (DCs) play an important role in identification and description of discourse relations since they are the most apparent pointers to discourse structuring on the surface, both for humans and machines. Whether a given expression is a DC or not always depends on the particular context. Some connectives are typical for “connective” relations (e.g., *protože – because, však – however*), some of them become DCs in certain contexts only (*jinak –otherwise*, *podobně – similarly*, *naproti tomu – on the contrary* [lit. *opposite this*], etc.). DCs are represented by different part-of-speech classes. According to the POS tagging scenario used for the PDT, discourse connectives are represented by the following PoS categories.[10]

a) coordinating conjunctions: *a* (*and*)*; ale* (*but*)*; však* (*but*)*; nebo* (*or*)*; proto* (*therefore*)*...*

b) subordinating conjunctions: *ačkoliv* (*although*)*; když (when*)*; místo, aby* (*instead*) ...

[10] For a detailed PoS characteristics of discourse connectives in the PDT see Mladová (2008a, p. 58–62).

c) particles (including rhematizers): *ovšem (however), zkrátka (in short), dokonce (even), také (also), například (for example)…*

d) adverbs: *potom* (*then*), *následně* (*afterwards*), *stejně* (*equally/alike*), *současně* (*at the same time*), *tak* (*so*), *totiž* (roughly *because, since, actually*)*...*

e) certain prepositional phrasess (prepositions + pronouns): *kromě toho* (*apart from that*), *k tomu* (*in addition to this*), *naproti tomu* (*on the contrary*), *tím* (*by this*) *…*

f) other parts-of-speech – mainly in case of fixed compound connectives: *na jedné straně* (*on the one hand*), *stručně řečeno* (*in short*), *jinými slovy* (*in other words*)*…*

g) elements formed by letters or numbers expressing enumeration: *a), b), 1., 2.…*

h) two punctuation marks: colon and dash.

The procedure of connective annotation has been already described in Section 3.3.1. We highlight again that the final decision about the function of an exprssion as a connective was up to the annotator. Also, the annotators were free to mark more expressions as a connective of one relation, in which way they were able to capture many modified connectives (*právě proto – exactly because*; *pouze tehdy, pokud – only if* [lit. *only then, if*]) or connective concatenations (*přesto však – nevertheless* [lit. *yet nevertheless*]; *a stejně tak – as well as* [lit. *and equally so*]). However, this approach required great attention in order to distinguish whether a co-occurence of more connective expressions means that they signal a single discourse relation or more relations at once. The latter possibility is demonstrated by *pak ale* (*but then*) in Example (23) in which there are two separate relations between the same arguments indicated by the two connectives, respectively: the *ale*-connective signals an opposition and the *pak*-connective indicates temporal succession.

(23) *Ta* [G. Sabatiniová] *už ve Flushing Meadows před čtyřmi roky triumfovala, ale pak ustrnula a posledních 40 turnajů vyšla naprázdno bez titulu.*

She [G. Sabatini] *triumphed already in Flushing Meadows four years ago, but then she stalled and her last 40 tournaments resulted with no title.*

3.4.1 Connectives with a Referential Component

Two important criteria for our delimitation of the DC category are (i) that connectives cannot be morphologically inflected and (ii) they do not represent grammatical constituents of a sentence. An acknowledged exception detected in course of the annotations are some uses of the Czech relative pronoun *což* (roughly *which,* or *and this* in other than attributive usage). It can represent an intra-sentential connective with a *conjunctive* meaning (cf. Example (24)), even though it can be inflected as a regular pronoun. Moreover, it plays a role of a participant in the sentence structure.

(24) *Válka nás sjednocuje, což pro nás není přirozené.*
The war unites us, which is not natural for us.

Another partial exception (from the inflectibility criterion) are prepositional phrases combining a preposition with an (inflected) form of the demonstrative pronoun *ten* (*this/that*), e.g. *naproti tomu – on the contrary* [lit. despite that], cf. the point (e) above. These expressions are partly fixed and in the connective readings fully interchangeable with basic discourse connectives.

To distinguish between connective and non-connective uses of the mentioned connections containing pronouns, we have proposed to make use of pronominal coreference (Poláková et al. 2012a). If the pronoun (the referential part of the expression) refers to an abstract object, i.e., in our annotation scheme, to its realization by a verbal phrase (clause), it represents a discourse connective. And on the contrary, if the pronoun refers to a "mere" entity, we do not evaluate it as a connective. Compare the following examples from Poláková et al. (2012a): Example (25) exemplifies a connective reading ("*apart from operating*") and (26) a non-connective reading ("*along with the catalog*") of the given prepositional phrases.

(25) *Mövenpick provozuje několik desítek hotelů nejen v Evropě, ale i v Asii a Africe.* Kromě toho **je známý i jako obchodní a potravinářská firma.**

Mövenpick operates dozens of hotels not only in Europe but also in Asia and Africa. Apart from that**, it is known also as a business and food company.**

(26) *British Library vydala stručný katalog knih uvedené tematiky čítající přes šest set položek z majetku knihovny.* K tomu *lze na místě zakoupit dvě publikace o ruské avantgardní knize, vydané specialistkou Susan Comptonovou.*

The British Library has released a brief catalog of topic-related books containing over six hundred items from the library property. Along with that [lit. with that] *you can purchase on-site two publications about the Russian avant-garde literature, published by the specialist Susan Compton.*

3.5 Arguments in the PDiT 1.0

The definition and delimitation of discourse units (arguments) as basic segments entering discourse relations is not straightforward. As mentioned in the introductory chapter, the Prague annotation scenario shares the basic notion of a discourse argument with the PDTB, namely the concept of *abstract objects* (AO) by Asher (1993). In general, abstract objects can be seen as various propositions, i.e. assertions about some set of entities (events, states, situations, facts, beliefs, questions, etc.).

3.5.1 Syntactic Structure of Discourse Arguments

Several syntactic constructions can be interpreted as AOs. It is mostly clauses, but also their nominalizations, deictic expressions referring to previous explicit propositions, sequences of more sentences etc. This is the theoretical view. In annotation practice, the projects aimed on marking large datasets had to restrict the annotation of AOs to a manageable subset. Mostly, discourse units (abstract objects) represented by clauses with finite verbs and partially some infinitive and participial constructions are annotated.

In the PDiT 1.0, discourse arguments are expressed by the following structures. Each type of structure is exemplified by a corpus example, cf. (27) to (37) below.

- a single clause:
  - independent (either as a simple sentence between two final punctuation marks (27) or as a part of a compound sentence (28))
  - dependent (29)
- a combination of clauses/sentences
  - a compound sentence (27) or any of its clausal subparts (30)
  - a sequence of sentences (31)
- a structure with an infinitive with the functor PRED (32)
- an elliptical structure
  - with a contextual verb ellipsis (33)

  - with a grammatical verb ellipsis (34)[11]

- a parenthesis (35)[12]
- a group of expressions with a finite verb not corresponding to a clause / continuous subtree (36)[13]

An exception was made in case of list structures. The so-called hypertheme, or the title of the list was annotated as a discourse argument even with a non-verbal structure (37). The relation between the hypertheme of the list and all the list entries was by default treated as a special case of *specification.* The entries of the list were then only annotated if their structure contained either a governing verb form or a colon. (cf. Section 3.7.1).

(27) *Dvojjedinost tohoto problému naštěstí pochopili poslanci: o počtu regionů a jejich působnosti chtějí rozhodnout současně.* **Chléb se** tedy **bude lámat ve sněmovně.**

Luckily the parliament members have understood the duality of the problem: they intend to decide on the number of regions and the scope of their authority at the same time. **All will** then **be decided in the House of parliament.**

(28) *Několikadenní cesta* sice *něco stojí,* ale **zákazníci se o kvalitě produkce přesvědčí na vlastní oči.**

A few days' journey may have its costs but **the customers may check the production quality by their own eyes.**

(29) Jelikož **na generálního ředitele Bohemie bez policejních zkušeností byla uvalena vazba**, *do Bruselu asi nepojede.*

As **the director general of Bohemia without any police experience was taken into custody,** *he will probably not go to Brussels.*

[11] In the contextual ellipsis of the governing verb, the elided verb is reconstructible from the previous context. It is mainly, but not only, an ellipsis of the second predicate in a coordinated structure. In grammatical ellipsis of the governing verb, the verb cannot be reconstructed from the previous context. For a detailed analysis of discourse arguments with verb ellipses cf. Poláková et al. 2012b, pp. 55–58.

[12] We are aware of the fact that parentheses can be syntactically also represented by clauses/sentences or sets of clauses/sentences. But since their relation to the rest of the discourse can be quite loose and so it has some impact on discourse coherence, we list them here as a special category.

[13] The mismatches in correspondence of arguments to (sub)trees are mainly due to attribution. Cf. Section 4.2.2 for details.

(30) *Podle přesvědčení majitelů dosáhla prosperity* zejména proto, že **zaměstnává lidi, na které se může spolehnout.**

According to the conviction of the owners she achieved prosperity mainly because **she employs people that she can rely on**.

(31) *Velice špatná je situace většiny důchodců, kteří představují zhruba čtvrtinu obyvatelstva. Minimální starobní důchod je 7260 Ft, ale i průměrný důchod stěží přesáhne 10000 Ft.* Tato částka není o mnoho vyšší než úhrada za provoz středně velkého bytu v topné sezoně. A tak **dilema zaplatit činži, anebo se najíst se pro mnohé stalo realitou.**

The situation of most pensioners, who account for about a quarter of the population, is very bad. The minimal retirement pension is 7260 Ft, but even the average pension hardly exceeds 10000 Ft. This amount is not much higher than the payment for running a middle-sized apartment during the heating season. And so **the dilemma whether to pay the rent or to eat has for many become a reality.**

(32) *To je jasné, že bych byl radši, kdyby tady dosud stál zámek a ne tohle monstrum.* Ale **proč o tom stále uvažovat.**

It is obvious that I would prefer if there still was a castle and not this monster. But **why keep thinking about it.**

(33) Je šéfem mocné vojenské komise při ÚV KS Číny – ale *armádní špičky si prý od něj udržují odstup.* Stejně tak **pekingští byrokraté.**

He is the head of a powerful military board at the CPC Central Committee – but *military leaders allegedly keep a distance from him.* Just like **the bureaucrats from Beijing**.

(34) *Odpověď arogantní*, **odpověď** ovšem **věcná.**

An arrogant answer, but **a factual one**.

(35) *V době studentských protestů přišel o všechny funkce* (**radil** totiž **k umírněnému postupu vůči demonstrantům**) a octl se v domácím vězení.

In the time of students' protests he lost all his posts (as **he recommended moderate steps against the protesters**) and he ended up under house arrest.

(36) *U výrazně barevné tetováže* se dá předpokládat, že *barva není uložena příliš hluboko*, tedy **půjde relativně snadno odstranit**.

For strongly coloured tatoos it can be expected that *the colour did not penetrate too deep*, and so **it will be relatively easy to remove**.

(37) [Kniha] *Je rozčleněna do tří částí. V první se zabývá finančním a kapitálovým trhem a jejich nástroji. Druhá je věnována burzám - jejich systému, obchodování na nich - a cenným papírům. Třetí pak pojednává o historii burzovnictví a konkrétně popisuje významné světové burzy.*

[The book] *Is divided into three parts. In the first one it deals with the financial and capital markets and their tools. The second one is devoted to the stock markets – to their system, to trading on them – and to securities. The third one addresses the history of stock exchange and describes specifically the world's major stock markets.*

3.5.1.1 VP Coordinations

In the PDTB 2.0, coordinations of verbal phrases were only annotated for connectives other than conjunctive. So, an example like (38) would not have been annotated for the absence of the pronoun in the second clause. As Czech is a pro-drop language, a large portion of conjunctive relations would have been lost in this way. Therefore, Czech "VP phrases", constructions with missing pronouns, were fully annotated.

(38) *Vodoměry se po jisté době prověřují a cejchují.*

After a certain time, the water meters are examined and calibrated.

3.5.1.2 Nominalizations

In the first release of the PDiT, as noted in the beginning of this section, no verb nominalizations were annotated as discourse arguments. To distinguish such a nominalization in Czech is easy, cf. Example (39). The suffix "–ní" is a typical suffix for Czech deverbative nouns. To express the same content with a clause in this case is possible but quite unhandy (40). The best English translation then is with an *ing*-participle (41).

(39) *Před přistáním* [Prep+N] *si zapněte bezpečnostní pás.* LP

(40) *Před tím, než přistaneme* [Before we land], *si zapněte bezpečnostní pás.* LP

(41) *Before landing fasten your seatbelt.* LP

Ing-forms can be seen as a transit between a verbal (e.g. *arrive*) and a nominal (*arrival*) way of expression an AO in English. In the PDTB 2.0, these structures were annotated as discourse arguments. In Czech, the strong tendency to fully nominalize similar English participle constructions lead to disproportion in what can be annotated within the same approach across languages. The variety of syntactic realizations of the same AOs in different languages shows the difficulty of drawing the same line in practical understanding of a discourse argument.

3.5.2 The Minimality Principle

In accordance with the PDTB annotation approach, the extent of an argument in the PDiT respects the *minimality principle* (Prasad et al. 2007, p. 14) which says that a discourse argument includes only such an amount of information that is *minimally required* and at the same time *sufficient* to complete the semantics of the relation. Any other relevant (but not necessary) information is in the PDTB annotated as supplementary information.

For PDiT 1.0, the minimality principle is related mostly to the number of sentences (trees) included in a single argument. Dependent clauses (mainly the attributive ones) within one tree were mostly considered a part of the argument. Having removed an attributive clause from an argument must have been justified.

3.5.3 Naming of the Arguments

In the PDTB annotation, the notation of the two discourse arguments is motivated syntactically: the clause associated with the discourse connective is marked Argument 2 (Arg2), the other is argument is marked Argument 1 (Arg1).

In the PDiT 1.0, in contrast, **the arguments have been defined semantically.** Thus, for instance, in the relation of *reason – result*, the text span expressing the reason is always marked Arg2, and the text span expressing the result is always marked Arg1, regardless which one contains the connective or in which order they appear in the text. **An important annotation rule is that the discourse link (the arrow) always leads from Arg2 to Arg1.**

These two principles match in majority of cases. They match for intra-sentential discourse relations of governing and dependent clauses (where a connective in form of a subordinate conjunction typically occurs), for all the symmetric relations (e.g. *confrontation*, *equivalence,* cf. Section 2.4.2.4) in which the semantic properties of the two arguments are the same, and for the class of EXPANSION. In the latter two cases, the Argument 1 is simply the one more to the left in linear order in the Prague annotation scheme.

The comparison of these two principles of argument labeling shows that the PDTB approach has easier accessibility to the information where the connective is to be found (look in Arg2 of any relation) whereas the Praguian approach has to use the link to the ID of the connective node associated with the discourse arrow together with the direction of the arrow.

On the other hand, thank to the semantic labeling of the arguments (represented by the oriented discourse link) in the PDiT, the Prague repertoire of asymmetric discourse relations can be reduced by half compared to the PDTB without loss of information.

We demonstrate the differences in naming conventions of the arguments on Example (42) and its translation to English: the Czech sentences are annotated for the connective *proto* (*therefore*) with the *reason – result* relation, Argument 1 being the text span to the right (the result) and Argument 2 being to the left (the reason). The English equivalent of these sentences would be in the PDTB-style annotated for *therefore* with "CONTINGENCY:Cause:Result" relation, with Arg1 to the left, Arg2 with the connective to the right. The situation is illustrated also with the different use of boldface (Arg2) and italics (Arg1):

(42) Naší výhodou je, že **v různorodých provozech Setuzy se surovina zpracuje beze zbytku**. *Nemáme* proto *potíže se získáváním trhu pro své výrobní odpady.*

Our advantage is that *in the various plants of Setuza the feedstock is processed completely.* **We have** therefore **no difficulties with gaining market for our production waste.**

3.6 Discourse Relations in the PDiT 1.0

The nature and the properties of discourse relations in general and in different frameworks have been described in the theoretical part of this thesis (Section 2.4.2). The present section is devoted specifically to discourse relations as annotated in the PDiT 1.0. It describes our treatment of discourse relations in connection with the settings of Prague treebanks and with the practical purposes of the annotation task itself. The core of this section is the semantic classification of discourse relations annotated in the Prague Discourse Treebank.

3.6.1 Hypotaxis and Parataxis

Two basic formal principles of grammatical arrangement of a sentence are hypotaxis (subordination) and parataxis (coordination). In syntax, and in particular in the European approaches to syntax, the notion of hypotaxis and parataxis are strongly connected to certain semantics. For instance, the meaning of condition is typically connected to the hypotactic form of expression, since the typical conjunctions with conditional meaning are subordinators.

In our analysis of discourse, we disregard these tendencies in formal arrangement of the sentence and we claim that the semantic types introduced for discourse mostly have both possibilities of expression. In our concept, discourse relations can be expressed hypotactically or paratactically within a single sentence, and further between individual sentences or larger text units. Example (43) demonstrates a conditional meaning expressed by paratactic means. Thus, the syntactic distinction hypotactic vs. paratactic does not play any role in the design of our semantic classification for discourse.

(43) **Posluchač musí přistoupit na pozici, že vše je dovoleno.** Potom *se pobaví a také pochopí, že drama znázorňuje ztrátu reálné komunikace.*

The listener has to accept the position that everything is permitted. Then *he enjoys [the play] and also understands that the drama symbolizes the loss of a real-life communication.*

3.6.2 Discourse Relations and their Semantic Types

The Prague set of semantic types for discourse relations was inspired by the tectogrammatical functors (Mikulová et al. 2006) and by the PDTB 2.0 sense tag hierarchy (Miltsakaki et al. 2008). The four main semantic classes, TEMPORAL, CONTINGENCY, COMPARISON and EXPANSION are identical to those in the

PDTB[14] but the hierarchy itself is only two-level, with a total of 22 relations. The third level of the Penn hierarchy is captured by the direction of the discourse arrow (cf. Section 2.4.2.4 on asymmetry). The annotators, in contrast to the PDTB annotation procedure, were not allowed to only assign the major class; they always had to decide for a single relation within one of the classes.

Within these four classes, the types of the relations partly differ from the PDTB types and go closer to Prague tectogrammatical functors. The discourse-semantic classification for the annotation in the PDiT 1.0, the Czech names of the relations and their annotation labels are given in Table 3.3. Appendix 1 then provides every relation type with an authentic PDiT example and its English translation.

Although we believe that the general interpretation of semantics in discourse is the same for Czech and English, and it is likely language-universal as for the main features, the repertoire of language means for expressing discourse functions is, on the other hand, largely language-specific. As such it can influence a fine-grained semantic classification (cf. Mladová et al. 2009). For the semantic classification of discourse relations in the PDiT, compared to the PDTB 2.0 label set, the CONTINGENCY class in the PDiT 1.0[15] was extended by the categories of *purpose* and *explication*, the CONTRAST (COMPARISON) class by *restrictive opposition* (which also includes the meaning of *exception*), *gradation* and *correction,* a category typical for the Czech connective *nýbrž*. *Correction* also includes the PDTB category *chosen alternative* (typically substitution with *instead*).

In the PDTB, four pragmatic meanings are distinguished and annotated: *pragmatic cause, condition, contrast* and *concession*. In the Prague scenario, three pragmatic senses were annotated. *Pragmatic concession* and *pragmatic contrast* were merged together as one group for the lack of reliable distinctive features. The PDiT annotation was initiated at the point where no further sub-classification of the pragmatic domain was carried out, as it is the case in some of the recently created discourse corpora. Hence, we did not distinguish between epistemic and speech-act readings of pragmatically used connectives (as discussed above in Section 2.4.2.1). In the annotated data of PDiT 1.0, however, both these readings can be found and the pragmatic “f”-labels can be revisited at any time.

We are well aware of the fact that the proposed semantic classification for discourse annotation in the PDiT is not the only one possible. Our decisions resulting in its present shape were motivated practically, by the annotation task

[14] With one naming exception: the COMPARISON class is referred to as CONTRAST class in the Prague scheme.

[15] There were no adjustments of the semantic classification towards the PDT 3.0.

English name	**Czech name**	**label**
TEMPORAL (ČASOVÉ vztahy)		
synchrony	současnost	synchr
asynchrony (precedence – succession)	nesoučasnost	preced
CONTINGENCY (KAUZÁLNÍ vztahy)		
reason – result	příčina – důsledek	reason
pragmatic reason – result	nepravá příčina – důsledek	f_reason
explication	explikace	explicat
condition	podmínka	cond
pragmatic condition	nepravá podmínka	f_cond
purpose	účel	purp
CONTRAST (KONTRASTIVNÍ vztahy)		
confrontation	konfrontace	confr
opposition	opozice	opp
restrictive opposition	restriktivní opozice	restr
pragmatic contrast	nepravý kontrast	f_opp
concession	přípustka	conc
correction	rektifikace	corr
gradation	gradace	grad
EXPANSION (NAVAZOVACÍ vztahy)		
conjunction	konjunkce	conj
conjunctive alternative	konjunktivní alternativa	conjalt
disjunctive alternative	disjunktivní alternativa	disjalt
instantiation	exemplifikace	exempl
specification	specifikace	spec
equivalence	ekvivalence	equiv
generalization	generalizace	gener

Table 3.3: Semantic types of discourse relations in PDiT 1.0 and PDT 3.0

and by the nature of our data. Only the annotation itself, and, even more so, an application of the classification on a different type of data (e.g. spoken) can test its adequacy. In principle, there are two types of difficulties: if a distinction between some of the semantic types is notoriously difficult, it is either due to inexactness or incomprehensibility of the relation definition or to a gap in the scheme, or there is simply no single correct solution. In authentic data, sometimes more than a single interpretation is acceptable, because the way we write and speak is often underspecified or unclear.

3.6.3 Discourse Relations and Valency

At the beginning of the project, when we compared the tectogrammatical functors against the discourse sense labels in the PDTB 2.0, we made the following observation: according to the concept of verb valency in FGD (cf. Section 2.2.1), none of the functors for obligatory modifications of verbs (or actants, with the function to complete the valency frame of the verb: actor, patiens, effect, addressee, origin, and marginally other, like location) ever functions also as a discourse semantic label. In other words, verbal valency relations and discourse relations are mutually exclusive. In the following examples, a relation of a dependent subject clause (44), object clause (45), locative/directional clause (46) to the main clause – if these sentence constituents are obligatory modifications of the main verb, are never perceived as discourse relations.

(44) *Není dosti pravděpodobné,* (*že by parlament přijal vyrovnaný rozpočet na věky věků...*) ACTOR

It is quite unlikely (*that the parliament would adopt a balanced budget for ever and ever...*) ACTOR

(45) *Dá se očekávat,* (*že dotyční soudci budou mít problémy*). PATIENS

One can expect (*that the concerned judges will be in trouble*). PATIENS

(46) *Zdá se, že lidé pochopili jednu staronovou pravdu, že kam nechodí smích,* (*tam chodí lékař*). DIR3 = direction ”where to”

It seems that people have come to understand one old-new truth that where there is no laughter, (*there has to be a doctor*). DIR3 [Lit. Where there comes no laughter...]

3.6.4 Secondary Relations

In some cases, and primarily in the class of temporal relations, two different relations can be recognized between the same two arguments. Either two relations are signaled by a single connective (*jakmile* – roughly *when, once, as soon as*; *dokud* – roughly *as long as, until, while)*, cf. Example (47) with *synchrony* + *confrontation*, and (48) with *condition* + *asynchrony*) or a connective co-occurs with another one (*pak ale* – lit. *then but; nejdřív ovšem* – lit. *first but*; *a zároveň* – *and at the same time*, cf. Example (49). In the first case, only the interpretation perceived as stronger was annotated. For the latter case, we have established a standardized annotator's comment *second_rel*. In the trees, we have marked the relation considered to be primary; the secondary relation is mentioned in the comment in the standardized form "second_rel RELATION_TYPE connective". In this preliminary fashion, it is present in the PDiT 1.0 release.

(47) *Zatímco Jelcin jedná v Tokiu, Ruská federace se atomizuje.*

While Yeltsin is at a meeting in Tokyo, the Russian Federation is falling apart.

(48) *Nemocnice přitom často neposkytne pomoc, dokud nemá potvrzení o solventnosti.*

The hospital often does not provide any assistance until it has a certificate of solvency.

(49) *Ta* [G. Sabatiniová] *už ve Flushing Meadows před čtyřmi roky triumfovala, ale pak ustrnula a posledních 40 turnajů vyšla naprázdno bez titulu.*

She triumphed already in Flushing Meadows four years ago, but then she stalled and her last 40 tournaments resulted with no title.

For co-occurrences of *a* (*and*) with a temporal connective, the temporal relation was taken to be primary (stronger), and in this case we did not mark any relation (*conjunction*) as a secondary relation in the comment. For co-occurrences of a contrastive connective with a temporal one, on the contrary, the contrastive meaning was taken to be primary. Here, the temporal connective was mostly assigned together with the contrastive connective to the contrastive relation: *ale pak* (*but then*), *ale současně* (*but simultaneously*), *zároveň však* (*but at the same time*). Only if the context suggested also a strong temporal meaning, we commented on the existence of a secondary temporal relation.

Comments on secondary relations between the same arguments realized by two different connectives have been processed separately later, by creating full discourse annotation. It is included, together with other updates, in the next release – the PDT 3.0.

3.6.5 Semantic Classification of Discourse Relations in the PDiT 1.0

So far, we have characterized the annotation of discourse relations and their semantic classification in the PDiT 1.0 in general. In this section, we offer a detailed description of each of the individual 22 types of relations in our discourse taxonomy. This characteristics combines instructions for the annotators as they were introduced in the annotation guidelines (Poláková et al. 2012b) and, in general features, description resulting from our own annotation experience.
Description of each discourse relation type contains the following information:

- The name of the relation
- Definition of the relation and its short characteristics
- Naming of the arguments (which one is Arg1 and Arg2)[16]
- Typical connectives in Czech (as they were documented in the PDiT 1.0)
- Real-data example and its annotation

Where applicable, we draw a comparison of our treatment of a given relation to its treatment in the PDTB 2.0. Where necessary, we illustrate the description of the annotation with a figure.

3.6.5.1 Temporal Relations

The basic semantics of temporal relations is described as A & B, where A is valid and B is valid; within this group, the two arguments are temporally related. On the level of sentential analysis, most of the relations that express certain temporal characteristics are already reflected within the tectogrammatical annotation. The tectogrammatical representation uses nine semantically differentiated temporal functors, which express various time points or periods (answering questions such as *from when, how long, how often, until when, since when* etc., cf. Mikulová et al. 2005, pp. 452–474).

In the PDiT 1.0 annotation, we mark, in accordance with the PDTB 2.0, two temporal relations: **asynchrony** (or precedence – succession) and **synchrony** (or simultaneity). Connectives of temporal relations belong according to their PoS characteristics either to temporal subordinating conjunctions (e.g. *až – when;*

[16] We emphasize here, again, an important annotation principle: the discourse link (the arrow) always leads from Arg2 to Arg1. In this way, the different nature of some discourse arguments is marked, together with the information about their order in the text.

než – before, until; dokud – as long as, until), or to temporal adverbs. For the temporal adverbs, however, it is not straightforward to distinguish which expressions really function as connectives and which do not. We have addressed this issue earlier (Mladová 2008a, pp. 96–98), pointing to the different functions of what Hoffmannová (1986) calls time and space "text pointers"[17]. She claims that these pointers express, with a different proportionality, two functions: (i) a concrete reference to a certain time (space) of an event; (ii) a cohesive function, a reference to temporal (spatial) placement in another, sometimes even a distant text unit (1986, p. 161). Hence, we did not annotate temporal expressions with a reference of the first type or, in other words, with strictly adverbial meaning: *yesterday, next week, daily, till Monday* etc. Also, temporal expressions close to what Jakobson calls "shifters" (Jakobson 1971, p. 130) such as *teď* (*now*), *dnes* (*today*), etc., which exophorically refer to the moment of speech and do not refer to the temporality of another proposition in the discourse, have not been annotated. Finally, constructions comparing two situations/events in two temporal settings with expressions such as *dosud* (*so far*)... *nyní* (*now*) or *dříve* (*earlier*)... *dnes* (*today*) have not been marked. Their meaning is not primarily temporal; their function is rather contrastive, cf. Example (50). They have been already annotated within the annotation of the topic-focus articulation (the so-called contrastive topic, the tag *c* in the tectogrammatical attribute *tfa* assigned to the nodes representing the temporal expressions). Thus, such constructions may be easily accessed automatically; we did not assign any confrontational or temporal meanings to them in the discourse annotation.

The context for Example (50): An inexpensive café for public is to be opened in the private spaces of the Chamber of Deputies. The security is not very happy about it.

(50) *Dosud totiž ozbrojení muži hlídali jen veřejnosti nepřístupné prostory. Nyní se budou přímo pod pracovnami poslanců pohybovat desítky cizích lidí.*

So far, the armed men have guarded only the off limits spaces. Now tens of strangers will be walking directly under the offices of politicians.

Even with these three criteria for determining temporal connectives for our annotation (in short: (i) no exact time references like *on Monday*, (ii) no shifters and (iii) no doubled temporals in contrastive positions), the annotators hesitated over some other expressions. Finally, we distinguished three groups of temporal expressions which function always/sometimes/never as discourse connectives

[17] *textové orientátory*

as a lead for the annotators. The interpretation of some of these expressions is moreover dependent on their context.

- Connectives in the majority of cases are: *potom, pak, poté, posléze, vzápětí, následně* (all synonyms for *then*, *thereafter*).
- Mostly connectives: *mezitím* (*meanwhile*), *dále* (*further*).
- Mostly not connectives (context dependent): *už* (*already*), *ještě* (*still*), *okamžitě* (*immediately*), *tehdy* (*at that time*), *nakonec* (*in the end*), *později* (*later*), *konečně* (*finally*), *nadále* (*hereafter*), *dosud* (*so far*), *opět* (*again*), *znovu* (*again*), *zatím* (*for now*).

Asynchrony (Precedence – Succession)

The relation of asynchrony has two realizations: either the order of the arguments in the text corresponds to their progress in time, cf. Example (51), or they are ordered against their temporal succession, Example (52). These two realizations of *asynchrony* are distinguished in our annotation by the direction of the discourse arrow, cf. Figures 3.4 and 3.5 corresponding to Examples (51) and (52), respectively. The Arg1 of *asynchrony* is always the proposition happening later in time.

(51) **Veškerý vliv nynějšího předsedy ČSSD vyšel v tu chvíli naprázdno** a posléze *zklamal i jeho pokus výsledky pražského sjezdu anulovat.*

All the influence of the current ČSSD chairman proved fruitless at that moment and later *his attempt to invalidate the results of the Prague congress failed, too.*

(52) *Štaidl s pomocí detektivní agentury vypátral zmizelou zpěvačku teprve po dvou týdnech.* Předtím **mu anonym telefonicky sdělil, že byla unesena.**

With a help of a detective agency, Štaidl only tracked down the disappeared singer after two weeks. Before that, **an anonymous call informed him that she had been kidnapped.**

Synchrony (Simultaneity)

In the *synchrony* relation, both propositions are happening simultaneously. Arg1 is by default the first argument in the linear order, cf. Example (53).

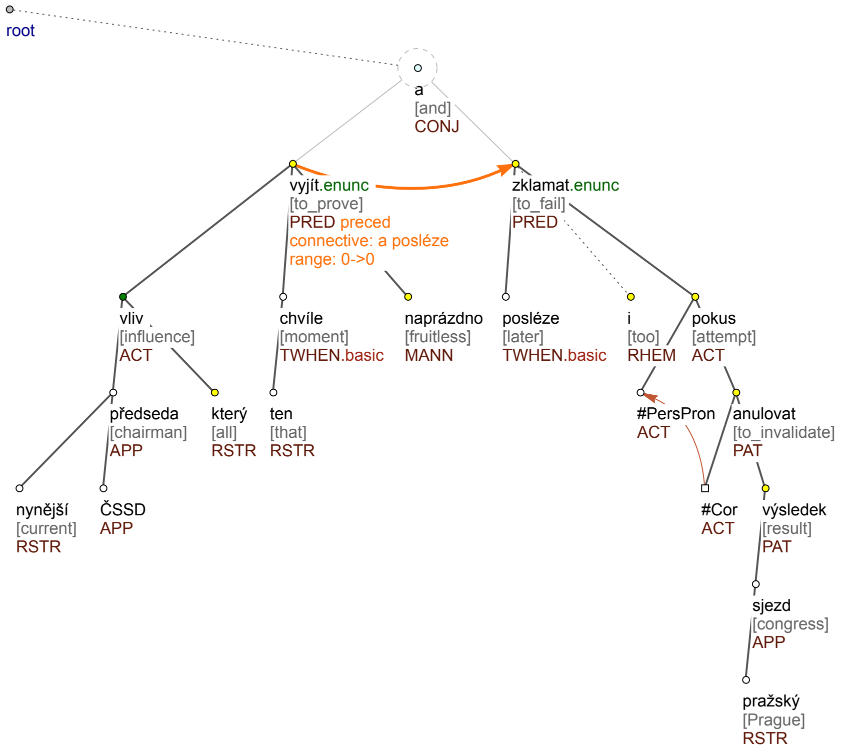


Figure 3.4: Asynchrony: Arg2 -> Arg1 order

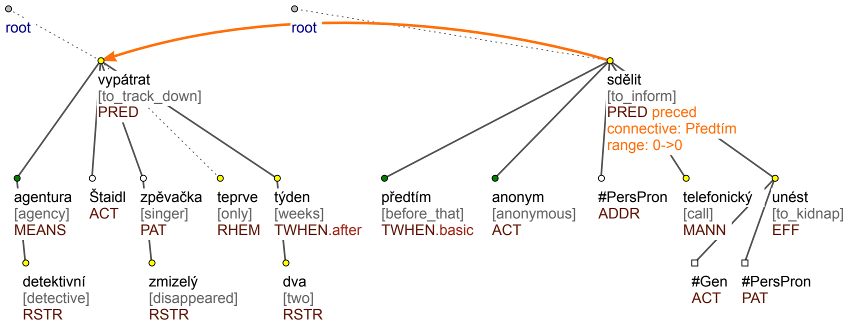


Figure 3.5: Asynchrony: Arg1 -> Arg2 order

(53) *Město postihla krize a nezaměstnanost.* Zároveň **začala nová éra svobodných celních zón.**

The city was affected by crisis and unemployment. At the same time, **a new era of free customs zones started.**

Some of the typical connectives expressing simultaneity require a closer attention: In a given context, they may express either temporal synchrony or, in a figurative meaning, they are means of structuring the text itself (not of the temporal structuring of the actual events) – cf. Example (54):

(54) P. Dvorský zahájí program áriemi od B. Smetany a A. Dvořáka. *K této literatuře se hlásím jako k vlastní, řekl Dvorský.* **Zároveň připomněl, že v Čechách se mu vždy dostávalo velké pozornosti.**

P. Dvorský will start the program with arias by B. Smetana and A. Dvořák. *I accept this literature as my own, Dvorský said.* **He** also [lit. *at the same time*] **noted that, in Bohemia, he had always received a great attention.**

3.6.5.2 Contingency Relations

The contingency class contains causal discourse relations in a broad sense. The basic semantics of the class may be expressed as A -> B, A implies B or A is (causally) related to B. Partially, this class includes the relation of *concession* for its causal component. However, for the annotation purposes, and also in accordance with the scheme of the PDTB and similar projects, we classify the concessive meaning as a relation from the CONTRAST class, for its contrastive component.

The annotation scenario of PDiT 1.0 contains four semantic and two pragmatic contingency relations: **reason – result, explication, condition, purpose; pragmatic reason – result** and **pragmatic condition.**

Reason – Result

Reason – result is a very frequent discourse relation, and, at least as our annotation statistics confirm (cf. Section 4.1 below), it is the most common way of expressing causality in the text. The linguistic distinction in syntax suggested by some of the Czech grammars (e.g. Daneš et al. 1987, p. 480; Svoboda 1956, p. 3) between a spontaneous, objective, natural **cause** (in Czech *příčina*) of events and an intended **reason** (*důvod*) for somebody's thinking/saying/doing something is not made in the PDiT, both are treated as the same demonstration of semantic causality.[18]

Similarly to *precedence – succession*, the *reason – result* relation is usually realized "in both directions", the order of arguments is arbitrary. In the PDiT annotation, the proposition expressing a reason is always considered to be Arg2. Example (55) demonstrates the order reason -> result with the connective *Proto* (*Therefore*) and Example (56) the order result -> reason with the connective represented by a colon.

(55) **Pivo, o jehož názvu by se mělo rozhodnout v průběhu tohoto týdne, je podle jeho slov vhodné zejména po tělesné námaze.** Proto *bude ve sklenicích o obsahu 0.25 litru nabízeno například ve fitnesscentrech a na plovárnách.*

The beer, the name of which should be decided in the course of this week is, according to his words, suitable especially after physical exercise. Therefore, *it will be offered in 0.25 liter glasses for example in fitness centers and at swimming pools.*

(56) Edvard Beneš byl tématem natolik kontroverzním, že přivedl do varu i nejserióznější historiky. **Není jim co závidět**: *Beneš patří mezi ty kultovní osobnosti, kterých si vážíme tím méně, čím více se o nich dovídáme*.

Edvard Beneš was a subject of so much controversy that he got inflamed even the most serious historians. **They are not to be envied**: *Beneš is one of those iconic figures we cherish the less, the more we learn about them.*

Pragmatic Reason – Result

The relation of *pragmatic reason – result* is annotated in such cases where causality does not hold between the two propositions themselves. In order to understand the causal relation the recipient must infer some content unexpressed by the author, from the context or from the recipient's world knowledge (cf. Section 2.4.2.1 on the different sources of coherence). The argument containing a pragmatic reason, analogically with the "semantic" relation of *reason – result*, is marked as Arg2.

In Example (57) below, the writer infers that the evaluation of the Prime minister by the public involves also evaluation of his responsibilities for others

[18] It is a future task to rethink the *reason – result* category and to introduce a distiction between a cause (příčina) and reason (důvod). We feel this task also needs a deeper insight into the pragmatic factors behind some of these relations.

from the fact that whenever the trust in others declined, the trust in the Prime minister declined also. Here, the causal relation lies between the author's assumption/claim and the support/evidence he gives for it. An easy paraphrase with a common causality marker here sounds: *I claim that A, because I know B.* This type of relation is treated as *Pragmatic cause: justification* in the PDTB, as *Explanation-argumentative* in RST, and as *epistemic Explanation* in TüBa-D/Z (Versley and Gastel 2013).

(57) **Při posuzování premiéra Klause bere veřejnost patrně v úvahu i jeho odpovědnost za činnost celého kabinetu, případně jednotlivých resortů.** *Dlouhodobé výsledky STEM* totiž *ukazují, že vždy, když klesala důvěra ve vládu, klesala i důvěra v premiéra.*

When evaluating the Prime Minister Klaus, the public apparently takes into account his responsibility for the activities of the whole Cabinet or the individual departments. *Long-time results of the STEM agency show* [totiž = roughly *as a matter of fact*] *that whenever the trust in government declined, the confidence in the prime minister dropped, too.*

Condition[19]

Within a sentence, the conditional relation is usually expressed hypotactically between clauses, namely by a clear and limited repertoire of subordinate conjunctions (*jestliže, kdyby, když, -li, pokud*, etc. (roughly *if, when*). Much more rarely, it is expressed asyndetically or by means of coordinating linking elements, modal verbs, the interrogative and the imperative verb mode. For the practical annotation, Arg2 is by default the proposition expressing condition, Arg1 the result of the condition. Example (58), mentioned earlier under (43), represents a less typical conditional relation: it holds between sentences and it is indicated by the connective *potom* (*then*).

(58) **Posluchač musí přistoupit na pozici, že vše je dovoleno.** Potom *se pobaví a také pochopí, že drama znázorňuje ztrátu reálné komunikace.*

[19] We are aware that the naming of the relation – *condition* means two things: the semantic type (meaning) of the relation and the nature of one of its arguments. In fact, the relation should be named *condition – result of the condition* in order to be consistent with other relations names (*reason – result*). The same is the situation for other relations like *pragmatic condition*, *purpose*, *concession.* Yet, for simplification, we use the shorter names for these relations.

The listener has to accept the position that everything is permitted. *Then he enjoys [the play] and also understands that the drama symbolizes the loss of a real-life communication.*

Pragmatic Condition

Similarly as in other pragmatic relations, the conditional meaning in a pragmatic condition does not hold between the two propositions themselves. The two propositions are causally unrelated; the validity of one is not determined by the other. Typically, in case of pragmatic condition, one proposition expresses a specific point of view, under which the other proposition is uttered. Further, this category includes some rare cases with a typically conditional connective with no conditional meaning at all and where no other clear semantic relation can be indicated. On the other hand, some Czech conditional connectives can regularly express the meaning of confrontation – those are not annotated as pragmatic conditions (according to their form) but as confrontations (according to their meaning). In annotations, similarly as in the case of the semantic condition, the proposition expressing the pragmatic condition is marked Arg2. Example (59) offers one of the most common structures annotated as pragmatic condition, a structure with speaker's evaluation of some fact (point of view).

(59) Jestliže **chcete slyšet můj postoj k rozhodnutí poroty**, *je to neslýchaný projev neúcty k práci druhého.*

If **you want to hear my attitude towards the jury's decision**, *it is an outrageous sign of disrespect for the work of others.*

Purpose

In the discourse relation of *purpose*, Arg2 expresses the purpose for which Arg1 is carried out. Purpose in Czech is primarily expressed within a sentence, as a clause element (so far irrelevant for our annotation), or as a dependent clause, cf. Example (60). Two independent sentences connected with the meaning of purpose are typically in the relation of *reason – result* with an additional intentional component expressed mainly by the verb *chtít* (*to want*) or synonymous expressions, cf. Example (61). For the *purpose* relation in Czech, we have not documented any inter-sentential connective in our data. Therefore, it would be possible not to consider *purpose* a type of a discourse semantic relation (as in the Penn Discourse Treebank 2.0). In some of the follow-up discourse annotation projects, however, a goal/ purpose relation was newly introduced (Hindi – Kolachina et al. 2012; Italian – Tonelli et al. 2010; French – Danlos et al. 2012).

For the annotation of PDiT, we also decided to introduce this relation, mainly in order to observe its inter- and intra-sentential distributions.

The relation of pragmatic purpose is not introduced in the PDiT 1.0, since the pragmatic ("false") purpose structures[20] are always interpretable as other semantic relations; cf. Example (62) which can be interpreted as *conjunction.*

(60) *Odcizené věci si vojáci uložili do svých skříněk* s tím, že **si je odvezou do civilu.**

The soldiers have stored the stolen things into their lockers in order to **take them with them into the civilian life.**

(61) *Marie pravidelně cvičí. Chce zhubnout.* LP

Mary works out regularly. She wants to lose weight.

(62) *Českou republiku opustí zítra,* aby **pokračovala do Rakouska, Moldávie a Zakavkazska a do Moskvy.**

She will leave the Czech Republic tomorrow in order to **continue to Austria, Moldova and the Caucasus and to Moscow.**

Explication

The main motivation for introducing the explicative relation in the PDiT annotation is a small group of (frequently used) Czech connectives that do not have their exact counterparts in English (*totiž, vždyť, přece*). They can be translated roughly as *as a matter of fact, actually, after all, yet.* According to Czech linguistic handbooks, they primarily relate "two propositional contents where the second one brings a clarification, explanation of the first one, important or necessary for its full understanding. The contents are, however, semantically independent of each other, there is no semantic relation of reason – result between them. Therefore, we speak about the simply explicative relation." (Grepl and Karlík 1986, p. 372).[21]

In this sense, the correlation between *cause/reason – result*[22] on the one side and *explication* on the other could be illustrated by the following scheme (Figure 3.6):

[20] in the Czech terminology: "nepravé věty účelové", cf. also Jínová et al. (2013).

[21] "dva propoziční obsahy, z nichž druhý přináší objasnění, vysvětlení prvního, důležité nebo nutné k jeho plnému a správnému pochopení. Obsahy však nejsou na sobě závislé sémanticky, není mezi nimi významový vztah příčinně-následkový. Mluvíme proto o poměru prostě vysvětlovacím."

[22] for the distinction between *cause* and *reason* cf. the subsection on *reason – result* (3.6.5.2)

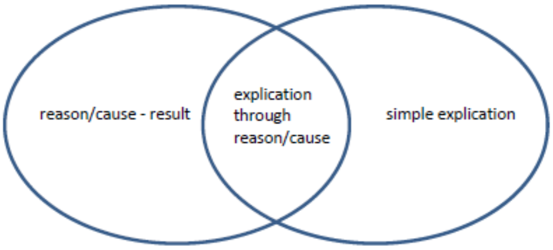


Figure 3.6: The relation of causal and explicative meanings

Accordingly, the relation of explication in the PDiT is annotated in those cases which express a non-causal explication of the content of Arg1 through the Arg2 (i.e. explaining without giving causal evidence). Example (63) shows a non-causal (simple) explication, which would be annotated as *explication* in the PDiT, whereas Example (64) demonstrates a causal explication, which would be annotated as *reason – result*. Example (65) then offers an authentic instance of the explication relation from the PDiT 1.0.

(63) *Hráli dobře, dali totiž pět gólů.* LP

They played well, as they scored five goals.

(64) *Hráli dobře, hodně totiž trénovali.* LP

They played well, they have trained a lot.[23]

(65) *Včerejší porada ministrů o státním rozpočtu na rok 1995 dopadla víc než dobře.* **Václav Klaus ani Ivan Kočárník** totiž **nenašli v Kramářově vile nikoho, kdo by se s nimi chtěl prát o ideu vyrovnaného rozpočtu.**

Yesterday's meeting of the ministers concerning the state budget for 1995 ended better than well. [Lit. As a matter of fact], **neither Václav Klaus nor Ivan Kočárník found anyone in the Kramář villa who would want to oppose them about the idea of a balanced budget.**

[23] Still, this kind of distinction has proven to be difficult to make in some situations. Deeper insight into the problem is offered in the analysis of M. Rysová (2010).

3.6.5.3 Contrastive Relations

The basic semantics of contrastive relations may be described as A & B, where the contents of the propositions A and B are different, dissimilar, contradictory or compared. The class of contrastive discourse relations in the PDiT 1.0 contains six semantic and one pragmatic relation. They roughly correspond to four of the PDT 2.5 syntactic relations that were marked previously within the tectogrammatical annotation: adversative (ADVS), confrontational (CONTR – subordinating, CONFR – coordinating) relations, concession (CNCS) and also gradation (GRAD). All these semantic categories were documented not only intra-sententially, but also between sentences and larger text units. For discourse annotation in the PDiT, they are marked as **opposition**, **confrontation**, **concession,** and **gradation.** In addition to these, new discourse relations have been established within the contrastive class: **correction** (or replacement), **restrictive opposition** (including exception), and **pragmatic contrast**. Their introduction was possible due to the fact that their formal and semantic properties can be easily described and distinguished by the annotators, compare the description of respective relations below in this section.

The semantic categorization of contrastive discourse relations has become in this way finer than the semantic categorization of contrast within a sentence (on the tectogrammatical layer). The annotation guidelines therefore instructed the annotators to further specify the tectogrammatical ADVS functor in intra-sentential relations (compound sentences), in cases where a "finer" distinction in meaning could be identified. Figure (3.7) demonstrates such a subclassification of the ADVS functor in the sentence in Example (66): a discourse link with the semantics of *correction* is annotated. In the second, automated step of annotation, all the remaining (not subcategorized) ADVS labels were extracted and converted into discourse relations with the semantic type *opposition.*

(66) [Opera Mozart] *Neprovozuje moderní hudební divadlo*, nýbrž **degraduje Mozartovu hudbu na pouhý kulisový doprovod k mnohdy samoúčelným jevištním skopičinkám.**

Lit: [Opera Mozart] *It-does not perform modern musical theater, but it-degrades Mozart's music to a mere stage set accompaniment to often purposeless stage foolery.*

It does not perform modern musical theater, **it** rather **degrades Mozart's music to a mere stage set accompaniment to often purposeless stage foolery.**

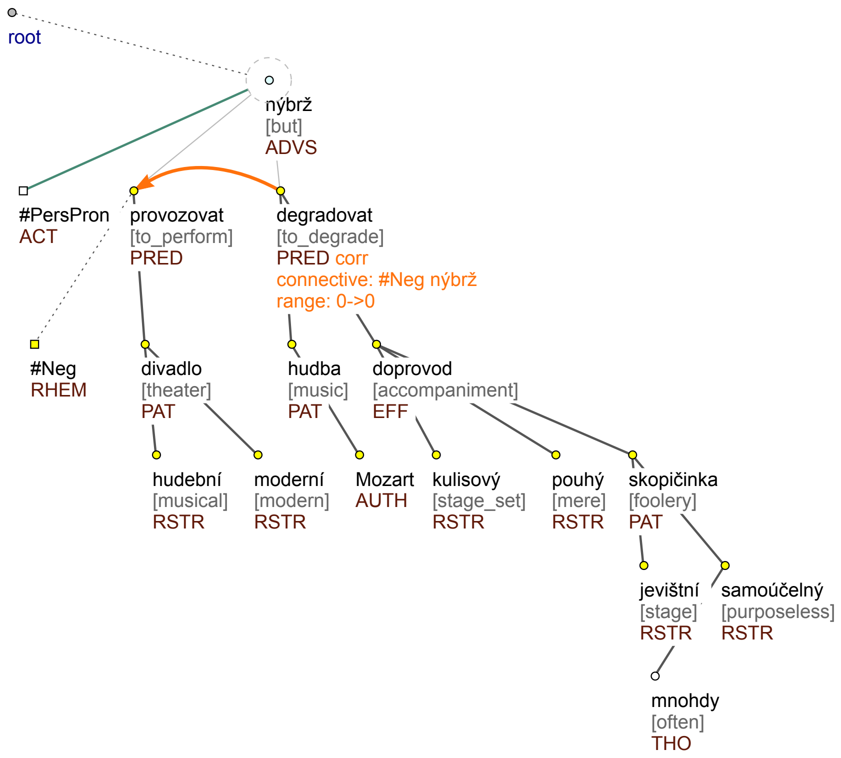


Figure 3.7: Subclassification of the ADVS functor (discourse type *corr*)

Confrontation

The relation of *confrontation*, within a sentence as well as between higher text units, indicates that two phenomena, situations, etc., have two different properties or a different degree of a single property. A simple scheme of *confrontation* is: "Component A has the property X, while component B has the property Y", where components A and B are from a certain set (e.g. people) and properties X and Y are somehow related – they are often two opposite poles on one scale. By convention, the first argument in linear order of the text is Arg1. The relation of *confrontation* is usually represented by the subordinating conjunction *zatímco* (*while*) or the coordinating conjunctions *kdežto* (*whereas*) and *ale* (but), and some adverbs like *naopak* (*on the contrary*), cf. Example (67).

(67) *Stejně dobře vykročila i Radka Bobková, jež vyřadila domácí Poovou 3:6, 7:5, 7:6.* **Nedařilo se** naopak **Ludmile Richterové, jíž vystavila stop ve třech sadách 3:6, 6:2, 4:6 další domácí tenistka Werdelová.**

Similarly well started also Radka Bobková who knocked out the domestic player Po 3:6, 7:5, 7:6. On the contrary**, things did not go so well for Ludmila Richterová who was eliminated in three sets 3:6, 6:2, 4:6 by another domestic tennis player Werdel.**

Opposition

The discourse relation of *opposition* expresses contrast or contradiction of two facts in a broad sense. Intra-sententially, it corresponds to the tectogrammatical adversative relation (ADVS). Unlike other relations in this group which have some additional semantic feature added to their basic contrastive meaning and therefore can be treated as more specific, *opposition* remains the basic way of expressing contrast. The basic connective for opposition in Czech is *ale* (*but*), cf. Example (68). Also here, the first argument in linear order is by convention Arg1.

(68) *Lidé chtějí platit jen to, co skutečně spotřebovali.* **Ještě dlouho tomu tak** ale **patrně nebude.**

People want to only pay for what they really have consumed. But **it apparently won't be this way for a long time yet.**

Pragmatic opposition

The relation of *pragmatic opposition*, similarly as *pragmatic reason – result* and *pragmatic condition*, does not express an opposition of the argument contents themselves. Usually, it expresses a relation between one argument content on one side and an unexpressed content (a presupposition or an inference) on the other; the form (as well as the connective) is adversative but the meaning is not clearly interpretable as opposition. This category contains both pragmatic concession and pragmatic opposition; these two types are not further distinguished due to the opacity of their meaning. We may also say that the arguments of this relation are vaguely contrastively connected or their meaning is, due to a large degree of inference, "at a great distance". By convention, the first argument in linear order of the text is Arg1.

(69) *Podle vedoucího výroby Miloše Přiklopila má Seba rozpracovanou celou řadu zakázek.* **Zákazníci** však **vyvíjejí velký tlak na snižování cen tkanin.**

According to the production manager Miloš Přiklopil, the Seba company has a range of factory orders in process. **The customers,** however, **exert great pressure on lowering the prices of fabrics.**

In Example (69), the connective expression *však* (*however*) does not express an opposition of the contents of the two sentences. Rather, it relates two inferences behind the stated facts: having a lot of orders implicates a major income to the company, but pressure to produce at a lower price can cause a decrease of the income.[24] The contrastive relation between the two sentences is perceived as a difficult one, "at a long distance". In other words, less coherent because of a higher "inferential load placed upon the hearer" (Grosz et al. 1995, p. 7), cf. also Section 2.4.2.1.

Restrictive Opposition

Restrictive opposition is a relation in which the content of one argument (the second argument in linear order) expresses a limitation of or an exception to the validity of the content of the other argument (the first one in linear order). The argument expressing the limitation or the exception is marked Arg2. On the tectogrammatical level, this discourse relation corresponds to the RESTR functor.

Some cases of restrictive opposition contain an implicit conditional meaning. In the annotation, we preferred to treat such cases as restrictive oppositions if the significance of the restriction or the exception prevails over the conditional interpretation, cf. Example (70).

(70) *Každá krajina má svou krásu.* Jenom **ji musíte umět vidět.**

Every landscape has its beauty. Only **you must be able to see it.**

The basic connectives of *restrictive opposition* are the connectives of opposition (*ale – but*, *však – however*) and some restrictive focusing particles, which are called rhematizers in the tectogrammatical annotation (e.g. *jen, jenom, pouze*, etc., all meaning *only*).

Concession

The semantics of discourse *concession*, similarly as the syntactic notion of concession, contains a causal component and a contrastive component. A concession is established with a denial in one of the discourse arguments of a certain expectation associated with the other argument. The causal component, the implication in an expectation (cf. *it is raining* implies: *people mostly do not go out*) is being denied while establishing the contrastive component of the concession (cf. *yet I will go out*). The nature of expectations in concessive relations influences the degree of coherence (comprehensibility) of the relation, compare

[24] The analysis of Example (69) is taken over from our article with Š. Zikánová et al., to appear in 2015.

the analysis of expectations in concessions mentioned in Section (2.4.2.1) above. Concessive meaning is primarily expressed intra-sententially and hypotactically (by governing and subordinate clause connection and the typical concessive connectives *přestože, ačkoli, i když* – all meaning *even though, although*) but it can be also expressed paratactically, intra- or inter-sententially, with help of contrastive coordinating connectives (cf. Example (71)) and discourse adverbials. In the annotations, the argument expressing the denial of an expectation is marked Arg1 (*yet I will go out*), the other argument (*it is raining*; in intra-sentential relations represented often by a dependent concessive clause) is marked Arg2.

(71) **Zdálo by se, že pirátské zboží zmizí z trhu.** Ale *po krátkém období paniky se překupníci a prodejci rychle vracejí k původní praxi.*

It would seem that the pirate goods would disappear from the market. But *after a brief period of panic, the traffickers and the dealers are quickly returning to the original practice.*

Correction

In the relation of *correction*, the content of the second argument in linear order corrects, replaces or substitutes the content of the first argument. One of the arguments is always (at least implicitly) negated – in the vast majority of cases the first one. A typical connective of correction is *nýbrž* (roughly *but, rather*)[25]; the negation morpheme associated with the verb is treated as a part of the connective. A typical pattern for the semantics of *correction* can be represented as: *Ne A, nýbrž B* (*Not A, but rather B*) and *Místo aby A, B.* (*Instead of A, B*).

The negation particle alone can also function as a connective of *correction* and it has been annotated as such, cf. Example (66) from the beginning of this section, here repeated as (72):

(72) [Opera Mozart] *Neprovozuje moderní hudební divadlo*, nýbrž **degraduje Mozartovu hudbu na pouhý kulisový doprovod k mnohdy samoúčelným jevištním skopičinkám.**

It does not *perform modern musical theater*, **it** rather **degrades Mozart's music to a mere stage set accompaniment to often purposeless stage foolery.**

In the tectogrammatical representation, most cases of correction within a single tree are marked with the ADVS functor but this relation occurs sometimes also in structures marked with the CONJ functor. These cases have been re-annotated.

[25] *nýbrž* in Czech corresponds the closest to the German conjunction *sondern.*

From the semantic point of view, the relation of *correction* also includes intrasentential structures with the tectogrammatical functor SUBS (substitution, replacement). However, these cases have been not re-annotated; they were directly extracted from the tectogrammatical annotation.

In the annotation, the negated or replaced argument is marked Arg1, i.e. in most cases, it is the first argument in linear order. In the following case (73), however, the typical order of the arguments is inverted:

(73) *Chytrý bankéř si klienty přece vytváří,* a ne ***se jich zbavuje.***

A smart banker rather creates his clients and **he does** not **get rid of them.**

Gradation

The discourse relation of *gradation* corresponds to the relation of gradation within a sentence (with the tectogrammatical functor GRAD). It compares a different degree of one property or two different actions, cf. Example (74). In some cases, it can be difficult to distinguish a *gradation* from a pure *conjunction.* In the annotation of PDiT, we mark only indisputably gradational connections. Gradation may also express a subjective judgement of events by the author and as such can be treated as pragmatic, cf. Example (75). In the annotations, Arg1 always expresses a lower degree of the property; the order of the arguments is arbitrary. Typical connectives of *gradation* are *navíc* (*moreover, what is more*)*; dokonce* (*even*); *nejen – ale i* (*not only – but also*) etc.

(74) *Letos se již zdálo, že počáteční nadšení místních radních pro tuto akci vychladlo.* **Organizátoři** dokonce **uvažovali o přemístění sympozia do Českých Budějovic.**

This year it already seemed that the initial enthusiasm of local councilors for this action had faded. **The organizers** even **considered relocating the symposium to České Budějovice.**

(75) *Pink Floyd pozdravili publikum, nadšeně reagující zejména na starší písničky, v průběhu koncertu několika českými větami.* **Ještě potěšitelnější** však **pod deštivým pražským nebem byla perfektní práce zvukařů.**

The Pink Floyd greeted the audience, which was responding enthusiastically especially to the older songs, with several Czech sentences during the concert.

But **even more heartwarming was the perfect job of sound engineers under the rainy skies of Prague.**[26]

3.6.5.4 Relations of Expansion

The basic semantics of the expansion class may be described as A & B, where B typically elaborates in a specific way on the content of A. The class of these elaborative discourse relations in the PDiT 1.0 contains seven semantic relations: **conjunction, instantiation, specification, equivalence, generalization, conjunctive alternative and disjunctive alternative**.

The nature of the expansion class, or the so-called elaborative relations, is different from the previous three classes; it is less often motivated syntactically (in terms of predicate verb modification, dependency and governing), it rather relates to the composition possibilities of a text – in the sense of set/subset membership (as in Kehler 2002, p. 18). In this respect, relations from the expansion class mainly determine how the content of an utterance elaborates on the content of either the preceding utterance or the whole previous section: whether it expands the content, brings a summary, gives examples, etc. These relations are also often present among larger text units such as paragraphs.

In the tectogrammatical analysis of a sentence, discourse relations in the expansion class correspond approximately to the tectogrammatical relations of conjunction (with the functor CONJ), disjunction (DISJ) and apposition (APPS). The realizations of these tectogrammatical relations between clauses of a sentence (not between noun phrases etc.) have been extracted and used for discourse annotation.

Apposition, from the perspective of possible text structuring means that are already captured in the sentence analysis, is a purely descriptive notion with no semantic information. It only indicates the content parallelism of the connected neighboring propositions. For this reason, all clausal (verb-containing) appositions (with the functor APPS) were re-annotated as some type of relation from the expansion class, usually as *specification*, *generalization* or *equivalence*. Figure (3.8) and Example (76) display an apposition of two clauses, annotated as *specification* with a colon as the connective.

(76) *Spisovatelovo umění se nezapře*: **málokomu se podaří vtěsnat tolik nenávisti a lži do jedné věty.**

The writer's art can not be denied: **Very few manage to squeeze such an amount of hatred and lies into one sentence.**

[26] In similar structures, even the comparative category in adjective forms can be regarded as having a connective function.

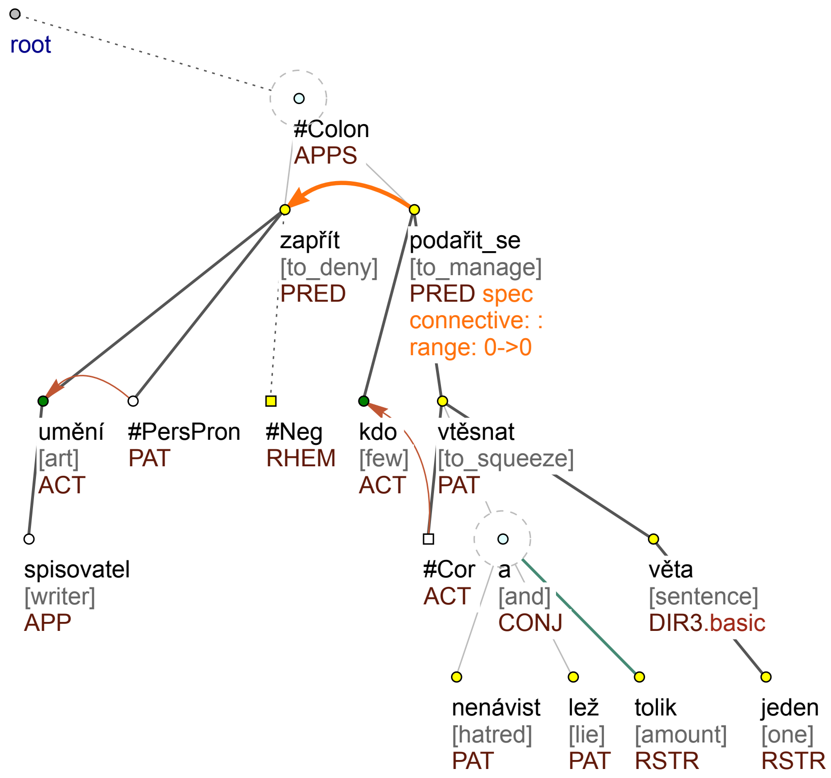


Figure 3.8: A structure with apposition (discourse type *spec*)

Conjunction

In the PDTB, the discourse relation of *conjunction* is defined negatively: “the situation described in Arg2 provides additional, discourse new, information that is related to the situation described in Arg1, but is not related to Arg1 in any of the ways described for other types of EXPANSION .” (Prasad et al. 2007, p. 37). Since this semantic relation is broad and partly vague, we stick with this definition. By convention, the first argument in linear order of the text is marked Arg1.

In the tectogrammatical representation, the conjunction marked with the CONJ functor and its clausal members correspond to intra-sentential discourse *conjunction*. However, in discourse annotation, there are some details to be treated separately:

(i) All clauses within one graphic sentence are always connected into a tree graph. If there is no semantic connection between the clauses, they are linked by a mere technical type of coordination with the CONJ functor. This means that compound sentences with the CONJ functor had to be manually checked for the actual existence of a true semantic relation (vs. a purely technical solution).

(ii) The Czech relative expressions *což* (roughly *which*), *přičemž* (*and, at the same time*), *čímž* (*thereby*), etc., are annotated primarily as conjunctions in the tectogrammatical analysis. Although CONJ is used appropriately in most cases, it was necessary to check these cases manually for a possible different relation.

(iii) With the conjunction *a* (*and*), if standing separately, we abandoned the implicit meaning of temporal succession. Only in cases of strong temporality with no temporal marker at all, where the whole construction could be easily transformed into a temporal one, the annotators were instructed to add a comment or to mark a secondary relation (cf. Section 3.6.4). If the conjunction *and* appeared together with another connective, the relation indicated by the other connective was annotated in the vast majority of cases: e.g., *a pak* (*and then*) (asynchrony), *a tedy* (*and so*), *a tak* (*and so*) (reason – result, explication etc.).

Typical connectives for the *conjunction* relation are *a* (*and*); *také* (*also*); *což* (roughly *which*); *rovněž* (*also*); *dále* (*further*) etc. An instance of a less typical connective (*kromě toho – in addition*, lit. *except that*) is given in Example (77) mentioned earlier under (25).

(77) *Mövenpick provozuje několik desítek hotelů nejen v Evropě, ale i v Asii a Africe.* <u>Kromě toho</u> **je známý i jako obchodní a potravinářská firma.**

Mövenpick operates dozens of hotels not only in Europe but also in Asia and Africa. <u>Apart from that</u>, **it is known also as a business and food company.**

Instantiation

In the relation of *instantiation* (also referred to as exemplification), the first argument contains a set (e.g. activities, behaviors, etc.) and the second selects its subset as an example. It is important that the two sets are not identical; the example represents a selection of the total, i.e. it is not that a single set is viewed from different perspectives. The Arg1 represents the more general proposition, the Arg2 the example.

Typical connectives are *například, třeba* (*for exampl*e), cf. Example (78). In the tectogrammatical representation, these expressions are usually evaluated as rhematizers (the RHEM functor). However, they can also function as connectives: if they open two positions for two discourse arguments, they participate in discourse composition (cf. Mladová 2008b).

(78) *Každá pověřená poradna spravuje agendu žadatelů o adopci v rámci větších územních celků.* Například **naše poradna v Kolíně působí ve dvanácti okresech středních Čech.**

Each authorized advisory office administers an agenda of adoption applicants within larger areas. For example, **our advisory office in Kolín operates in twelve districts of central Bohemia.**

Specification

In the relation of *specification*, the second argument expresses a detail or other specific information for the statement in the first argument. Like *instantiation*, it presents a subset with respect to the content of the first argument, but not an example.

Specification in the PDiT also contains the relation of a "hypertheme" (title) of a list structure and the group of all items in the list; cf. Section 3.7.1. Typical connectives are punctuation marks colon and dash; they are annotated the same way as regular connectives. In our experience (cf. Jínová et al. 2014), specification also occurs very often with no explicit connective. Such cases, however, have not been annotated yet. Within a single sentence, many appositions (the APPS functor) can be viewed as specifications, as stated earlier in this section: Figure 3.8, Example (76)). The Arg1 represents the more general statement, the Arg2 the more specific one.

(79) *V souladu se západními vzory je možná i omezená preference soukromého pojištění před sociálním pojištěním.* Konkrétně, **pokud si výdělečně činná osoba zaplatí dostatečně vysoké soukromé pojištění, bude se moci ze sociálního pojištění „vyvléknout“.**

In line with the western models, a limited preference of the private insurance over the social insurance is possible. Specifically, **if an employed person pays a high enough private insurance, they can "wriggle out" of the social insurance.**

Equivalence

The relation of *equivalence* combines two arguments in which the content of the propositions is the same, only expressed each time "in different words". The two arguments stay on the same level of generality, cf. Example (80). The second argument in linear order is neither more specific (*specification*) nor more general (*generalization*) in relation to the first one. These "restatement" relations (cf. the PDTB hierarchy) are quite similar: they introduce a claim and further elaborate on it in the following text unit. Often it is impossible to keep a clear boundary between these three relations in an authentic text, and the semantics can be determined only by the connective. The first argument in linear order is labeled Arg1.

Typical connectives of equivalence are *tedy* (*and so, which means*); *tak* (*so)* and *tj.* (*i.e*).

(80) Dnes nebo zítra se v dolní komoře polského parlamentu - v Sejmu - očekává hlasování, které bude mít vážné politické důsledky, *ať už dopadne jakkoliv,* tj. **bude-li zákon odmítnut či přijat.**

Today or tomorrow the lower chamber of the Polish Parliament – the Sejm – expects voting that will have serious political consequences *whatever the outcome will be,* i.e. **whether the law will be rejected or accepted**.

Generalization

The relation of *generalization* expresses generalization or summarization: the second argument in linear order contains a summary of the content of the first argument (cf. Example (81)) or, very often, it summarizes over multiple preceding propositions. The Arg1 designates the less general proposition, Arg2 the summarizing proposition. Typical connectives are mostly adverbial expressions such as: *stručně/kráce řečeno* (*shortly*); *tedy* (*so, thus*); *zkrátka* (*shortly, simply*); *prostě* (*simply*) etc.

(81) *Naše čtenářka, která by uzavřela životní pojištění na 20 let na pojistnou částku 100 tisíc s měsíčním pojistným 310 korun, by se mohla úrazově připojistit na dalších 100 tisíc za 32 korun měsíčně, zároveň by tím byla připojištěna i na úraz s trvalými následky na 200000.* **Ročně by** tedy **zaplatila na pojistném, včetně úrazového připojištění, 4104 korun.**

Our reader, who would take out a life insurance for 20 years for an insured sum of 100,000 CZK with a monthly fee of 310 CZK could take out also an accident insurance for an additional 100,000 CZK for 32 crowns a month, at the same time she would be insured against an injury with permanent damage for 200,000 CZK. Thus, **she would pay annually 4,104 crowns, including the accident insurance.**

Conjunctive alternative

Conjunctive alternative expresses a relation where the two arguments represent alternatives or options that may both hold at a given time. In Example (82), it is possible to interpret the relation between Arg1 and Arg2 as two alternatives that may but do not have to exclude each other. The first argument in linear order is labeled Arg1.

The conjunction *nebo* (*or*), which is typical for *conjunctive alternative*, can express the meaning of conjunction (the CONJ functor) in the tectogrammatical representation. All such cases connecting structures with finite verbs were left aside in the manual annotations. Subsequently, they were automatically extracted as instances *conjunctive alternative* in the semi-automatic phase of PDiT annotation.

(82) ...schopní lidé se dnes již věnují pouze své profesi, neboť *na amatérské působení mimo svou odbornost již nemají čas* nebo **se jim to prostě nevyplácí.**

...talented people today are dedicated only to their profession, as *they no longer have time for amateur activities outside their expertise* or **such activities just don't pay off.**

Disjunctive alternative

Disjunctive alternative expresses a relation where one argument excludes the other one, i.e. only one of the alternatives can hold at a given time, cf. Example (83). This is also the main difference between *conjunctive* and *disjunctive alternative.* The first argument in linear order is labeled Arg1. This relation basically corresponds to the DISJ functor in compound sentences in the tectogrammatical representation, but, in our view, in some cases, the DISJ functor also corresponds to *conjunctive alternative.* All instances of this functor connecting structures with finite verbs had to be manually checked for their actual discourse meaning.

(83) Proto je obzvlášť tristní poznání, že *vlády na krátící se termín blokace zákona o bankrotu zřejmě jednoduše zapomněly.* Nebo **mu nevěnovaly dostatečnou pozornost.**

It is particularly sad to realize that *the governments have apparently simply forgotten about the deadline for blocking the bankruptcy law.* Or **they just did not pay enough attention to it.**

3.7 Other Annotated Phenomena

3.7.1 List structures

Unlike in the PDTB annotation scheme, a list structure in the Praguian approach does not have a semantic label in the sense hierarchy. It is annotated as a separate phenomenon for two reasons: First, in this type of structure, every item of the list is related to the preceding item AND to the (facultative) introductory statement for the whole list, if present. Thus, the nature of a list structure is not strictly binary in the sense of our discourse relation definition. Second, we treat list structures as more or less compositional, formal phenomena in text organizing, with no semantic content. In our viewpoint, there is only a *specification* relation between the hypertheme (introductory statement) and the set of list items. So, in the annotation, the list structure is marked with a special attribute (the attribute *discourse/type* has the value "list", cf. Section 3.2.2 on data representation) and, facultatively, there might be also a *specification* relation. If so, the hypertheme of a list is the only exception in the notion of a discourse argument: for our annotation purposes, it does not have to include a finite verb. Also, there does not have to be an explicit connective connecting the hypertheme and the list items. Relaxing these two general annotation rules helps us preserve linguistic information about list structures in the annotation.

An example of a list structure with a hypertheme and two list entries is given in (84). The first sentence is the hypertheme; the connectives are the numbers *1.* and *2.*

(84) *K tomu, aby zaměstnavatel pracovníkovi za škodu opravdu odpovídal, musí být splněny tyto podmínky:* [hypertheme]
1. Zaměstnanci musí vzniknout škoda, tj. musí dojít k urcitému snížení hodnot jeho majetku (v některých případech mu vzniká i právo na náhradu ušlého zisku).
2. Zaměstnavatel nebo jiná fyzická ci právnická osoba, která jedná jeho jménem, musí porušit své právní povinnosti.

For the employer to be truly responsible for the suffered damage of an employee, the following conditions must be satisfied: [hypertheme]
1. The employee must suffer a damage, i.e. there must be some reduction in the value of his or her property (in some cases, there is also entitlement to loss compensation).
2. The employer, or other natural or legal person acting on his behalf, must violate their legal obligations.

3.7.2 Headings

Headings are annotated within two different attributes in the PDiT 1.0 and the PDT 3.0. In the PDiT 1.0, the attribute *is_heading* may have the values "0" or "1". If there is no value in the attribute, "0" is assumed. The value "1" is assigned to the root of the (sub)tree that represents a heading or a subheading in a text. In the PDT 3.0, the attribute *discourse_special* is introduced, with three possible values: "heading" (for marking headings and titles of the corpus texts), "caption" (for marking captions of photos, tables and charts) and "metatext" (for metatext information occurred by mistake during corpus compilation), cf. Section 3.2.2.

Headings and subheadings are annotated without distinction. Authors' names, their abbreviations, the location and the source of the article or other information regarding the text have not been marked in any way, as they are, in contrast to headings, a rather optional piece of information.

3.7.3 Captions

Clauses or sentences that in the original newspaper served as captions for figures, photos, charts or tables represent a specific type of information in the PDiT texts. They are often single-sentenced documents or they are attached to the texts of longer articles, where they cause incoherence. For the PDiT 1.0 annotation, the annotators added a standardized comment (a preliminary fashion of marking several marginal text phenomena) "is_photo" to any such caption they could recognize while annotating. In the PDT 3.0, this information is included in the attribute *discourse_special*. Example (85) presents a caption of a photo.

(85) *Bývalého generála sovětského strategického letectva nezapře Džochar Dudajev vzorně salutující na slavnostní přehlídce uspořádané při příležitosti třetího výročí vyhlášení nezávislosti Čečenska na Rusku. Foto Reuter*

Dzhokhar Dudayev cannot deny being a former general of the Soviet Strategic Air, saluting perfectly at the festive parade organized on the occasion of the third anniversary of the declaration of independence of Chechnya from Russia. Photo Reuter

3.7.4 Collections

The PDiT data contain a certain number of documents consisting of a set of short unrelated texts, mostly collections of short news reports of one to five sentences each. Thus, the whole of the document is incoherent and as such worth marking and excluding from any further work with these data for coherence modeling. The annotators of the PDiT 1.0 were instructed to add a standardized comment *collection* to the t-root of the first sentence in such a document. Later, in the PDT 3.0 release, the annotation of collections became a part of the genre specification of the corpus texts. A document representing a collection of different texts is demonstrated by Example (86).

(86) *Krátce*
Návrhy britského premiéra J. Majora a jeho irského partnera J. Burtona na budoucí uspořádání Severního Irska získaly včera podporu britské vlády. Dokument se stane v příštích týdnech předmětem diskusí konstitučních severo irských politických stran.
Dvěma hlavními cíli české zahraniční politiky jsou členství v Evropské unii a Severoatlantické alianci, řekl včera český ministr zahraničí Josef Zieleniec ve výboru pro zahraniční věci a zahraniční obchod Poslanecké sněmovny kanadského parlamentu.
Dohodu o zastavení palby porušil další ozbrojený konflikt mezi armádou a povstaleckou organizací UNITA, ke kterému došlo u severoangolského města Uige.
Irácká vláda nadále v "děsivé" míře a "bez jakýchkoli známek zlepšení" pošlapává lidská práva, konstatuje zvláštní zpravodaj OSN pro Irák Max van der Stoel ve zprávě, která byla včera zveřejněna v ženevském sídle OSN.
Zatím nelze říci, kdy bude sestavena nová polská vláda, řekl po setkání představitelů polské vládní koalice, Polské lidové strany a Svazu demokratické levice koaliční kandidát na křeslo premiéra, maršálek Sejmu J. Oleksy.

Briefly
Yesterday, the proposals of the British Prime Minister J. Major and his Irish partner J. Burton on the future organization of Northern Ireland received the support of the British government. The document will be a point of discussions of constitutional Northern Irish political parties.
The two main goals of the Czech foreign policy are the membership in the European Union and in NATO, the Czech Minister of Foreign Affairs Josef Zieleniec said yesterday in the Committee of Foreign Affairs and Foreign Trade Chamber of Deputies of the Parliament of Canada.
Another armed conflict between the army and the rebel organization Unita, which occurred at north Angola city of Uige, broke the agreement on cease-fire.
The Iraqi government keeps trampling on human rights in an "appalling" extent and "without any signs of improvement", UN special reporter for Iraq, Max van der Stoel says in his report, which was published at the Geneva UN headquarters yesterday.

> *So far, it cannot be said when new Polish government would be formed, the coalition candidate for the seat of Prime Minister Marshal of the Sejm J. Oleksy said after a meeting of representatives of Polish government coalition, the Polish People's Party and the Democratic Left Alliance.*

3.7.5 Metatexts

Typesetting information of the newspaper that occurred by mistake most likely during the corpus compilation is marked as "metatext" in the PDiT 1.0 and the PDT 3.0. It is formulations like *konec podnadpisu* (*end of the subheading*); *text do rámečku* (*frame text*) etc.

3.8 Checks and Evaluation

3.8.1 Post-annotation Checks and Fixes

After the manual annotation of discourse relations in the PDiT 1.0 was finished, some checks turned up to be necessary, especially for relations whose nature revealed to be more complicated in real data than we had expected initially on the basis of linguistic handbooks. We have collected all instances of these relations (namely *specification*, *explication*, *generalization*, *exemplification* and *equivalence*) in our data and established more specific delimitations among them. Annotation of these relations was manually unified in the whole data. Also, some connectives required unification via post-annotation. Additionally, the part of the data which was annotated first (train-1) was fully re-annotated since we expected it might have suffered from the initial inexperience of the annotators.

Results of the automatic extraction of intra-sentential relations were checked randomly on several hundreds of examples. Corrections of the discrepancies were integrated in an automatic script (treatment of multiple connectives, multiple coordinations etc.). Only two situations required manual fixing: (i) Due to tree complexity, the automatic extraction failed in 23 cases of connective identification (opposed to 10,482 cases with correct identification); (ii) Solely manual treatment was necessary for constructions with a discourse-relevant clause dependent on a complex predicate structure with an infinitive or a noun phrase. In such cases, only semantics allowed to distinguish whether a given dependent clause is related to the whole governing structure or only to its governing infinitive or noun phrase (cf. Jínová et al. 2012).

3.8.2 Automatic Checking Procedures

During the manual annotation of discourse relations in the PDiT 1.0, we collected proposals on automatic checking procedures that allow either to directly find and correct errors of certain types or at least to suggest where an error probably occurs. Some of these checking scripts were created and implemented continuously with every part of processed data, other were used later on all data at once. The list of rules applied in the checking procedures follows:

Rules that are always valid:

- every relation (arrow) is provided with a connective – on condition that it is the type "discourse" (not "list");
- nodes from/to which discourse arrows lead are either complex nodes (*nodetype* = "complex") with the value of "v" in the grammateme *sempos* or roots of coordinate structures (*nodetype* = "coap") or they are quasi-complex (*nodetype* = "qcomplex") and have a t-lemma substitute "#EmpVerb". In other words, governing nodes of discourse arguments are represented either by verbs or by coordinating expressions;
- at least one arrow leads from each group/to each group;
- a group consists of fewer nodes than one tree (and does not form a subtree in the tree), or on the contrary more nodes than one tree;
- a list structure includes more than one item;
- if the attribute *start_range* is assigned to a node in the last tree of the document, it can only have values "0" or "group".

Rules with possible exceptions:

- the arrow of a "list" type usually has a connective;
- the arrow of a "discourse" type with a *discourse_type* "spec" usually has a connective;
- the attribute *is_heading* usually belongs to the effective root of the tree;
- every node with the PREC functor is either a connective or it is provided with an annotator's comment (exceptions are *tak* (*so*) and *pak* (*then*) in pairs like *jestliže – pak* (*if – then*); *pokud – tak* (*if – so*); *když – tak* (*when – so*) etc.);
- every file usually contains at least one attribute *is_heading* = "1".

3.8.3 Inter-Annotator Agreement Measurement

During the process of manual annotations of discourse relations, we have regularly measured the inter-annotator agreement (IAA). The method, the process and the results of this annotation evaluation described in this section is summarized according to Poláková et al. (2012b).

The whole volume of the PD(i)T data, i.e. 49,431 sentences, is divided into 10 sections – 8 of them serve as training data for automated processes (train-1 – train-8), the two remaining sections are development and evaluation test data (dtest and etest). For the annotation purposes, each section was further divided into five roughly equal parts (of approximately 1000 sentences) that were given to the annotators, respectively. In each of these parts, we have selected an overlap of approx. 200 sentences. This overlap was annotated in parallel by all annotators who worked on the given part. The annotators did not know which files were selected for the evaluation (altogether 2,084 sentences in 44 documents).

The IAA measurements in the PDiT naturally only concerned the manual part of the project, i.e. inter-sentential relations and only such intra-sentential relations that differed from the tectogrammatical analysis (cf. Section 3.1.3). Relations within a sentence were extracted mostly automatically from the tectogrammatical tree structures and so the IAA measurement was irrelevant here.

The annotators were considered to be in agreement in recognizing a discourse relation if there was a non-empty intersection in connectives they had marked for a relation. For evaluation of this agreement, we used F1 measure, a harmonic mean between recall and precision (standard evaluation measures for such a task). Agreement on discourse types (semantic labels) was measured on discourse relations recognized by both annotators and is evaluated in two ways – first by a simple proportion of relations to which the annotators assigned the same discourse type, and second by Cohen's kappa, which measures the agreement on discourse types "above chance". In addition to the regular IAA measurements throughout the whole project, we have carried out a complex cross-sectional measurement on all parts of the treebank. We compared data from the only two annotators who have annotated all sections of the treebank. The results of the measurements are presented in Table 3.4.

In Table 3.4, we can observe a slightly rising tendency in agreement on connectives (F1 measure) and the highest agreement on discourse types in the most recently annotated section. These results, in our opinion, reflect annotators' gradual acquisition of experience with the texts and also some enhancements of the annotation concept. Table 3.5 shows the average values of F1 measure, agreement on types and Cohen's kappa on types for all data annotated in parallel.

Measurement	F1	Agreement on types	Kappa on types
train-2	0.83	0.69	0.57
train-3	0.79	0.8	0.75
train-4	0.8	0.75	0.69
train-5	0.85	0.76	0.71
train-6	0.84	0.77	0.68
train-7	0.79	0.67	0.61
train-8	0.86	0.84	0.79
dtest	0.85	0.73	0.67
etest	0.83	0.72	0.68
train-1	0.84	0.91	0.88

Table 3.4: Inter-annotator agreement measured on the parallel data in all sections of the treebank

Measurement	F1	Agreement on types	Kappa on types
all parallel data	0.83	0.77	0.71

Table 3.5: Inter-annotator agreement measured on all parallel data

These results can be considered satisfactory for the given type of annotation. For example, the average agreement on types (0.77) corresponds to the results in the Penn Discourse Treebank 2.0, where the agreement on types on the second level of the sense hierarchy reached 0.8 (cf. Prasad et al. 2008). However, as stated earlier, we have measured the IAA for inter-sentential relations and only a few intra-sentential relations.

Agreement and disagreement on the class level is presented in Table 3.6. In the contingency table, the cells along the diagonal represent the number of cases in which the annotators agreed on the class, other cells represent all variants of confusion.

	contrast	**contingency**	**expansion**	**temporal**	**total**
contrast	137	2	5	1	145
contingency	1	49	5	0	55
expansion	4	8	60	3	75
temporal	0	1	1	7	9
total	142	60	71	11	284

Table 3.6: Contingency table of agreement on four major semantic classes: contrastive relations, contingency relations, expansion relations, temporal relations (between two most productive annotators, all manually annotated relations in the IAA sample of data)

In our opinion, this table demonstrates a fair annotation consistency in terms of general semantic classes. Within the individual classes (cf. Table 3.7), there is quite often a disagreement among the individual types of the CONTRAST class and among some of types from other classes (*explication* and *reason*). This information offered us a valuable feedback for checking and adjusting the interpretation of the given relations on the basis of real-text data.

	conc	**cond**	**confr**	**conj**	**corr**	**exempl**	**explic**	**gener**	**grad**	**opp**	**preced**	**reason**	**restr**	**spec**	**synchr**	**total**
conc	7	0	0	0	0	0	0	0	0	1	0	0	1	0	0	9
cond	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
confr	0	0	7	0	0	0	0	0	0	2	0	0	1	0	0	10
conj	1	0	2	38	0	0	0	0	0	0	0	1	0	1	2	45
corr	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	4
exempl	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	2
explic	0	0	0	0	0	0	2	0	0	0	0	4	0	0	0	6
gener	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
grad	0	0	0	1	0	0	0	0	2	0	0	0	0	0	0	3
opp	0	0	2	0	0	0	0	0	0	58	0	0	2	0	0	62
preced	0	1	0	0	0	0	0	0	0	0	3	0	0	0	0	4
reason	0	0	0	1	0	0	1	1	0	1	0	17	0	0	0	21
restr	0	0	0	0	1	0	0	0	0	2	0	0	2	0	0	5
spec	0	0	0	1	0	0	0	0	0	0	0	0	0	4	0	5
synchr	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
total	8	1	11	41	5	2	3	1	2	64	3	22	6	5	3	177

Table 3.7: Contingency table of the inter-annotator agreement on semantic types of discourse relations (between two most productive annotators, inter-sentential relations only)

4

Learning from Discourse Annotation

Having thoroughly described the annotation scheme, process and consistency evaluation of the PDiT 1.0 (and the PDT 3.0), in this chapter we provide some reflections on the annotated treebank(s), the methodological concept used and bring new insight into the possibilities of annotation extension. We focus in particular on three aspects: the quantitative analysis of the annotated discourse phenomena (4.1), the analysis of one of the most apparent source of annotation inconsistencies – syntax – discourse mismatches (4.2) and on implicit discourse relations (4.3).

4.1 Corpus Statistics

The present section offers basic statistics and distribution figures for the annotated discourse phenomena. Basic numbers are presented for both the first version of the data, the PDiT 1.0 (released in November 2012), and for the updated and extended annotations in the PDT 3.0 (released in December 2013).[1] More detailed figures, regarding mostly connectives and arguments, are presented for the latest version (PDT 3.0) only. The reported statistics refer to the full extent of the corpora, which is 49,431 sentences in both cases.[2] Smaller tables are presented directly in this section. A larger table, (4.8), can be found in the Appendix 2, but, for easy reading and for illustration, its first few lines are also presented within this section.

[1] The differences in annotation between the two versions of the treebank are described in Section 2.2.5 above.

[2] We are aware of the other option to leave a small part of the data unobserved as testing data for NLP purposes. Yet, the size of the PDT treebank is approximately the same as the size of the annotated PDTB 2.0 (49,208 sentences), which is why, in this section, we report numbers for the whole corpus, too. The situation is different in Section 4.2 where the reported figures refer to the 9/10 of the treebank.

Relations	PDiT 1.0	PDT 3.0
discourse relations (arrows)	20,542	20,556
inter-sentential relations	6,195	6,226
intra-sentential relations	14,347	14,330
all relations (including lists)	20,903	20,917

Table 4.1: Discourse relations figures in PDiT 1.0 and PDT 3.0

4.1.1 Relations

There are 20,556 discourse relations annotated in the PDT 3.0 (20,542 in the previous data version PDiT 1.0)[3], cf. Table 4.1. Among them, 30.3% (6,226) are inter-sentential and 69.7% (14,330) are intra-sentential. As explained earlier in Section 3.1.3, all inter-sentential relations were annotated manually whereas intra-sentential relations were annotated semi-automatically. The latter include manual annotation of relations treated in a different way on the tectogrammatical and on the discourse level (1,951 instances), hand-crafted rules for extraction of temporal relations (643 instances), and finally automatic mapping of tectogrammatical labels, argument spans and connectives for the remaining intra-sentential relations (11,736 instances). This means that 42.9% of all discourse relations were annotated manually and 57.1% were treated semi-automatically. However, numerous manual post-annotation checks (esp. of connectives of intra-sentential relations, of the categories of generalization, equivalence and explication) and manual annotations of phenomena other than discourse relations (list structures, headings, captions, etc.) have to be taken into consideration (cf. Section 3.8).

4.1.1.1 Semantic Types of the Relations

Distributions of semantic types of discourse relations are given in Table 4.2. There are 22 annotated categories.[4] Within the major four classes, 43 % of the relations belong to EXPANSION, 28.9% to CONTRAST, 22.9% to CONTINGENCY and only 5.2% are TEMPORAL relations. In the individual types, almost two thirds of the relations are represented by three most frequent relations: conjunction (36.5%, from the EXPANSION class), opposition (15.5%, from the CONTRAST

[3] Again, all these discourse relations are associated with explicit connectives

[4] A 23th category “other” is added for two relations not provided with a semantic label by mistake – we mention them here for sake of completeness.

class) and reason – result (12.8%, from the CONTINGENCY class). The three pragmatic categories are the least frequent. This may be due to the nature of the treebank texts (written journalistic style as a predominantly objective type of discourse) or because of the difficulty (and our initial inexperience) of setting the borderline between "semantic" and "pragmatic" in the annotations.

4.1.2 Connectives

There are altogether 20,709 tokens of discourse connectives annotated in the PDT 3.0 (20,693 in the previous data version PDiT 1.0). Among them, 20,461 are connectives in discourse relations and 248 are connectives in list structures (between the individual list entries, associated with a list-type arrow). According to the annotation principles, connectives in list structures are not obligatory. Also, specification relations which relate a list's hypertheme (introductory statement) with the set of the list items do not have to contain a connective. This is why the overall numbers for discourse relations (20,556, cf. Table 4.1) are higher than the overall numbers for connectives in these relations (20,461).[5]

Annotation of discourse connectives in the PDiT 1.0 and in the PDT 3.0 was a task partly based on annotator's own decisions; this fact should be reflected in the annotation evaluation. Contrary to the PDTB annotation procedure, there was no predefined list of expressions to be annotated. As described above in Section 3.3.1, the PDT annotators followed some basic criteria for identification of connectives in the text. The results of the PDT 3.0 annotation therefore mirror this partial freedom: we have the advantage to learn about what the annotators considered to be actually connective expressions and what could have been omitted otherwise, with a predefined set of connectives. On the other hand, such an annotation principle naturally leads to a large group of sparsely occurring connectives (co-occurrences of various connective words and punctuation marks in one connective token or rich modifications of connectives).[6] These are then more difficult to categorize linguistically, for instance, to set the number of basic types of connectives and their modifications (as distinguished in the PDTB annotation manual (Prasad et al. 2007). There are 791 connective types (different strings)

[5] There are altogether 95 relations without connectives in the PDT 3.0. 43 of them relate to the specification relations in list structures, but 52 are errors – the connectives are absent by mistake. These cases are already fixed in the current (so far unpublished) data version.

[6] We assume that, in accordance with the annotation rules, a multiword connective once annotated in the PDT 3.0 is monofunctional, that means, it represents only one of the discourse relations (in contrast to the study of Webber et al. (1999) on multiple connectives which are polyfunctional).

Relation type	**PDiT 1.0**	**% (PDiT)**	**PDT 3.0**	**% (PDT3)**
total	20,542	100	20,556	100
TEMPORAL	1,030	5.0	1,066	5.2
CONTINGENCY	4,690	22.8	4,701	22.9
CONTRAST	5,930	28.9	5,938	28.9
EXPANSION	8,890	43.3	8,849	43.0
conjunction	7551	36.8	7498	36.5
opposition	3209	15.6	3196	15.5
reason – result	2626	12.8	2632	12.8
condition	1369	6.7	1369	6.7
concession	878	4.3	880	4.3
asynchrony	808	3.9	840	4.1
confrontation	654	3.2	653	3.2
specification	627	3.1	630	3.1
correction	440	2.1	445	2.2
gradation	430	2.1	445	2.2
purpose	414	2.0	414	2.0
disjunctive alternative	270	1.3	272	1.3
restrictive opposition	269	1.3	269	1.3
explication	225	1.1	230	1.1
synchronous	222	1.1	226	1.1
exemplification	142	0.7	148	0.7
generalization	106	0.5	106	0.5
equivalence	104	0.5	105	0.5
conjunctive alternative	90	0.4	90	0.4
pragmatic contrast	50	0.2	50	0.2
pragmatic reason – result	40	0.2	40	0.2
pragmatic condition	16	0.1	16	0.1
other	2	0	2	0

Table 4.2: Distribution of semantic types of discourse relations in the PDiT 1.0 and the PDT 3.0

annotated, 425 of which only have a single occurrence, cf. Table 4.3. This part of annotations definitely offers new information about non-typical ways of connecting discourse units or about the ways of possible composition and co-occurrence of connectives in Czech. Examples of rarely used connectives (87) – (88) demonstrate this. However, it is beyond the scope of this thesis to linguistically classify such an amount of annotated material in a complex fashion.

Example (87) documents a rare multi-part connective *zčásti – zčásti* (*partly – partly*), which functions analogically as other similar Czech connectives, but it is usually not listed in Czech grammar books – neither as a discourse connective nor as a connecting phrase within a sentence:

(87) *Dnes, kdy byl zčásti omilostněn, zčásti bylo jeho trestní stíhání zastaveno, uvažuje Alexej Žák o svém návratu do Čech.*

Today, when he was partly pardoned, partly his prosecution was stopped, Alexej Žák is considering returning to the Czech Republic.

In Example (88), *nejen* (*not only*) occurs as a separate connective. In Czech, *nejen* is nevertheless obligatorily followed by an occurrence of *ale i /ale také* (*but also*) or by other similar expression, thus forming an obligatorily two-part connective. Its independent occurrence signals some specificity in the annotated text. In (88), the second, the “but also”-part of the connective is missing. The meaning of gradation must be read out of the context: Not only is the HZDS party ready to abolish the voucher privatization, it is also prepared to re-nationalize big Slovak companies. We assume that we can understand and accept the omission of the second part of the connective in this case, because we can actually find it already in the title. The title of the article itself contains both the second argument of the gradation relation, which is the main point of the article, and the “but also”-part of the connective, the Czech *i* (roughly *and even*). Without the title, understanding would need a more complex inferencing process. Such a way of text structuring is unusual and stylistically odd in Czech, yet, as we have demonstrated, it is possible.

(88) *HZDS je připraveno i na znárodňování*
Bratislava-
Na Slovensku včera oficiálně začala volební kampaň k předčasným parlamentním volbám, které se budou konat 30. září a 1. října. O přízeň více než tří milionů voličů se bude ucházet 18 seskupení. Mečiarovo HZDS, které ve shodě s průzkumy mínění ohlašuje svůj návrat k moci, je připraveno zrušit nejen kuponovou privatizaci.

Connectives	**PDiT 1.0**	**PDT 3.0**
all connectives	20,693	20,709
connectives in discourse relations	20,445	20,461
connectives in list structures	248	248
connective types (strings)[7]	752	791
single occurrences of a connective	398	425

Table 4.3: Annotation figures for discourse connectives in the PDiT 1.0 and the PDT 3.0

V případě, že současná slovenská vláda prodá do voleb zahraničním zájemcům akcie velkých slovenských společností, HZDS je po volbách může opět znárodnit, prohlásil podle deníku Lúč místopředseda HZDS Sergej Kozlík.

The HZDS is prepared even for nationalization
Bratislava-
In the Slovak Republic yesterday, the campaign for early parliamentary elections, that will take place on September 30th and October 1st, officially started. 18 political groups will try to win the favor of more than three million voters. Meciar's HZDS, which is, in line with opinion polls, announcing its return to power, is ready to abolish not only the voucher privatization. In case that before the elections the current Slovak government sells stocks of large Slovak companies to foreign bidders, the HZDS can nationalize them again after the elections, vice-chair of the HZDS, Sergej Kozlík, said according to the daily Lúč.

In sum, the rare instances of discourse connectives in the annotated data of the PDT have the potential to reveal information about connective use so far unaccounted for in the Czech linguistic literature.

A remark on alternative lexicalizations of connectives:
There is no more than 200 occurrences of alternative lexicalizations of connectives (the "altlex" category, e.g. *v případě, že* (*in case that*); *přes tyto skutečnosti* (*despite these facts*)) in the PDT 3.0, where they were only annotated in a preliminary fashion. Recently, after the PDT 3.0 release, their annotation was updated and most of them are now called secondary connectives (Rysová and Rysová 2014).

4.1.2.1 Frequency of Connectives

The following three tables present figures about frequency of connectives in both versions of the treebank. Table 4.4 presents frequencies for connectives in discourse relations (for 40 most frequent connectives). There are 32 connective types with the frequency higher than 100 and 53 types with the frequency higher than 50 in the PDT 3.0. Table 4.5 presents 20 most frequent connectives in list structures, the reported numbers are identical for the PDiT 1.0 and the PDT 3.0. In Table 4.6, frequencies of "non-word" connectives (punctuation marks and the negation morpheme: the prefix *ne-* of Czech verbal forms) are given. Here, the number in brackets represents the number of connectives where the "non-word" element only represents a subset of the connective, e.g. negation morpheme + *nýbrž* (roughly *not + but*). Semicolon and three dots, originally not intended to be annotated as connectives, got among the annotated punctuation marks in the few cases where the annotators tried to correct a sentence segmentation problem.

The two versions of the treebank differ only slightly in connective frequencies: for some frequent connectives, decrease in total occurrences is visible. This goes hand in hand with the manual revisions of some semantic types of the relations. For instance, some of the occurrences of the most frequent connective *a* (*and*) were reassessed as not a "mere" conjunction, but rather a part of a complex connective indicating a different semantic type.

4.1.2.2 Ambiguity of Connectives

Findings about ambiguity (or polysemy) of discourse connectives are presented only for the most recent data version, the PDT 3.0. Table 4.7 shows discourse connectives with the highest variation in the assignment of semantic types and the percentage of their prevalent semantic type. But, right the first row of Table 4.7 shows that the connective *a* (*and*), even though it has assigned 13 semantic types, is not that ambiguous: 98.73% of all occurrences belong to the prevalent category of conjunction. Also, three of the remaining semantic types of *a* have only a single occurrence. That is why we present Table 4.8, where the ambiguous connectives are sorted according to the decreasing entropy of the distribution of their semantic categories (full extent of Table 4.8 is given in Appendix 2). Entropy figures take into consideration both the number of different categories assigned to the connective and their proportion, thus better reflecting the uncertainty in predicting a semantic category for the given connective.[8] We computed the

[8] The entropy enumerates the uncertainty of choosing a semantic category for a given connective under the assumption that the only available information is the probability distribution of possible semantic categories for the connective, which we approximate by the distribu-

PDiT 1.0		**PDT 3.0**	
a	5,947	a	5,820
však	1,535	však	1,527
ale	1,286	ale	1,275
když	576	když	575
protože	525	protože	525
totiž	461	totiž	461
pokud	405	pokud	404
:	396	:	396
proto	380	proto	380
tedy	308	tedy	308
aby	304	aby	306
ovšem	295	pak	296
pak	287	ovšem	293
-li	249	-li	249
také	234	také	234
neboť	222	neboť	221
-	220	-	218
zatímco	204	zatímco	204
nebo	191	nebo	191
což	189	což	189
navíc	184	navíc	182
i když	183	přitom	181
přitom	181	i když	178
sice ale	170	sice ale	168
naopak	154	naopak	152
takže	149	takže	149
a tak	143	a tak	141
dále	118	a to	118
a to	118	dále	117
kdyby	116	kdyby	116
tak	111	tak	112
rovněž	108	rovněž	107
přesto	100	proto že	99
#neg ale	98	přesto	99
přestože	98	#neg ale	98
proto že	98	přestože	98
například	95	například	97
zároveň	93	zároveň	94
přičemž	90	a proto	86
či	86	či	86

Table 4.4: 40 most frequent connectives excluding connectives in list structures

Connective	Frequency
*	46
-	37
1 .	11
2 .	11
3 .	9
4 .	7
5 .	6
první	5
třetí	5
za prvé	5
6 .	4
a)	4
b)	4
za druhé	4
10 .	3
7 .	3
8 .	3
9 .	3
c)	3
d)	3

Table 4.5: 20 most frequent connectives in list structures in the PDT 3.0

Connective form	PDiT 1.0	PDT 3.0
negation morpheme	70 (329)	71 (332)
colon	396 (418)	396 (417)
dash	220 (259)	218 (260)
... (three dots)	6	6
semicolon	3 (5)	3 (5)

Table 4.6: “Non-word” connectives in the PDiT 1.0 and the PDT 3.0

entropy (H) only for connectives with more than 50 occurrences in total and list the semantic type distributions for 40 connectives with the highest entropy figures.

Sorting connectives according to the entropy indicates, for example, that the two connectives with the highest entropy values, *když* (*if, when, while, as, because* etc.) and *přitom* (*at the same time, nevertheless, although* etc.), even though they have three less semantic categories assigned than *a* (*and*), have a much higher prediction uncertainty than *a* (H = 2.28 for *když*; H = 2.26 for *přitom*, but H = 0.13 for *a*). The connective *a* (*and*) is only 46th in the chart, cf. Appendix 2).

Table 4.9 then lists non-ambiguous connectives with more than 15 occurrences (only one semantic category assigned, H = 0).

tion of the categories for the given connective annotated in the PDT 3.0 data. The entropy is defined as the negative of the logarithm of the distribution; it is measured in bits (if the base of the logarithm is 2) and its value can be equal to or greater than 0 (0 in case of no uncertainty).

Connective	No. of disc. types	Prevalent type (%)	
a	13	98.73	conj
-	12	49.54	spec
ale	10	84.47	opp
:	10	72.98	spec
když	10	41.74	cond
tedy	10	60.39	reason
tak	10	75.89	reason
přitom	10	42.54	conj
však	8	81.20	opp
ovšem	8	77.82	opp
kdy	8	30.77	reason
totiž	7	69.63	reason
a to	7	52.54	spec
jinak	7	40.91	disjalt
aby	6	94.77	purp
pak	6	64.19	preced
i když	6	89.89	conc
sice .. ale	6	88.10	opp
jestliže	6	87.95	cond
avšak	6	80.33	opp
aniž	6	41.18	opp
neboť	5	96.38	reason
což	5	96.30	conj
takže	5	91.95	reason
i	5	91.57	conj
přičemž	5	88.37	conj
jenže	5	76.39	opp
či	5	62.79	disjalt
nicméně	5	57.75	opp
a přitom	5	57.89	opp

Table 4.7: 30 most ambiguous connectives in the PDT 3.0 sorted by the number of different semantic types assigned

H (Entropy)	Connective	Semantic type	Occurrences	%
2.30	**když**	cond	240	41.74
		preced	116	20.17
		synchr	95	16.52
		spec	71	12.35
		reason	26	4.52
		conj	11	1.91
		conc	7	1.22
		explicat	5	0.87
		confr	3	0.52
		restr	1	0.17
2.26	**přitom**	conj	77	42.54
		conc	41	22.65
		opp	36	19.89
		confr	8	4.42
		grad	6	3.31
		restr	4	2.21
		spec	4	2.21
		synchr	3	1.66
		reason	1	0.55
		f_opp	1	0.55
2.12	**- (Dash)**	spec	108	49.54
		reason	39	17.89
		conj	37	16.97
		explicat	21	9.63
		equiv	3	1.38
		confr	2	0.92
		opp	2	0.92
		gener	2	0.92
		conjalt	1	0.46
		restr	1	0.46
		cond	1	0.46
		exempl	1	0.46

Table 4.8: Ambiguous connectives in the PDT 3.0 sorted according to the decreasing entropy of their semantic categories – first three connectives.

Connectives with non-ambiguous meanings[9]	**Translation (a possibility)**	**Discourse type**	**Frequency**
také	also	conj	234
a proto	and therefore	reason	86
#Neg - negation particle	#Neg	corr	71
ačkoli	although	conc	48
ačkoliv	although	conc	40
že	that	reason	34
pokud ... pak	if ... then	cond	32
poté	afterwards	preced	30
ač	although	conc	27
byť	although	conc	25
přesto že	lit. although (that)	conc	23
buď ... nebo	either ... or	disjalt	22
a tudíž	and so	reason	20
jelikož	because	reason	20
a dokonce	and even	grad	19
pokud ... tak	if ... then	cond	19
a ani	(neither) ... nor	conj	18
#Neg ... spíše	#Neg ... rather	corr	17
li ... pak	if ... then	cond	16
např.	e.g.	exempl	16
později	later	preced	16
tím že	lit. thus (that)	reason	16

Table 4.9: Non-ambiguous connectives in the PDT 3.0 with more than 15 occurrences

4.1.3 Arguments

Table 4.10 gives basic overview about the extent of the two arguments of inter-sentential discourse relations in the PDiT 1.0 and PDT 3.0. Intra-sentential relations are assumed to appear within the given sentence and are not included. The distribution of "smaller" and "larger" arguments shows that 89% of inter-sentential relations (5,568 in total) have two single sentences or their parts as both arguments (the values of *start/target range* are 0->0). For a correct interpretation of Table 4.10, cf. the data representation given in Section 3.2.2. Table 4.10 also shows cases where an argument does not correspond to a dependency (sub)tree (the value "group" of the attributes *start/target_range*): 60 relations (1%).

The values of the range attributes, however, do not say anything about the mutual position of the two arguments, i. e. about their (non)adjacency. With a PML-TQ query based on the id number of a t-tree, we found out that 1,163 relations have at least one extra sentence between their two arguments and so can be viewed as non-adjacent. More details to the extent and location of the arguments are given in the analysis in Section 4.2.2.

PDiT 1.0				PDT 3.0			
start_range		target_range	count	start_range		target_range	count
0	->	0	5532	0	->	0	5568
0	->	1	234	0	->	1	236
1	->	0	110	1	->	0	116
0	->	2	50	0	->	2	50
0	->	group	43	2	->	0	37
2	->	0	36	0	->	group	31
0	->	backward	31	0	->	backward	30
group	->	0	29	group	->	0	23
0	->	3	22	0	->	3	22
0	->	4	13	0	->	4	14
0	->	5	12	0	->	5	11
3	->	0	8	3	->	0	10
0	->	6	8	0	->	6	8
0	->	7	7	4	->	0	7
4	->	0	7	0	->	7	7
1	->	1	5	1	->	1	6
backward	->	0	5	backward	->	0	6
0	->	8	5	5	->	0	6
5	->	0	5	0	->	8	5
0	->	10	4	0	->	10	4
0	->	9	4	0	->	9	4
group	->	group	4	0	->	12	2
7	->	0	2	1	->	group	2
0	->	13	2	group	->	group	2
1	->	group	2	0	->	13	2
0	->	12	2	7	->	0	2
0	->	21	1	6	->	0	2
9	->	0	1	1	->	3	1
1	->	3	1	group	->	1	1
1	->	2	1	0	->	16	1
4	->	1	1	0	->	21	1
forward	->	0	1	3	->	2	1
1	->	4	1	1	->	4	1
group	->	forward	1	4	->	1	1
0	->	11	1	forward	->	0	1
6	->	0	1	0	->	11	1
0	->	16	1	0	->	51	1
3	->	2	1	9	->	0	1
0	->	51	1	1	->	2	1
				group	->	forward	1

Table 4.10: Argument extents in PDiT 1.0 and PDT 3.0

phenomenon	representation/way of counting	PDT 3.0
list structures	first item of the list	84
list entries	starting nodes of list arrows	445
all relations	all relations represented by arrows (discourse arrows + list arrows)	20,917[10]
headings	value "heading" of the attribute *discourse_special* of a t-node	4,187
captions	value "heading" of the attribute *discourse_special* of a t-node	242
metatexts	value "heading" of the attribute *discourse_special* of a t-node	53

Table 4.11: Figures for other annotated phenomena in the PDT 3.0

4.1.4 Other Annotated Phenomena

There are 84 list structures in both versions of the treebank, with together 445 list entries. Relations between individual entries within a list structure are not considered discourse relations in our annotation, only the relation between a hypertheme (title) of the list, if any present, and all its entries is considered a discourse relation (specification, cf. Section 3.7.1). That is why the total number of all relations (arrows) in the treebank does not equal to the sum of discourse relations and list structures. 84 discourse relations are present in lists. Also, we have annotated 4,187 article headings and subheadings, 242 captions of photos, charts and tables and 53 instances of metatext (text not belonging to the article content), cf. Table 4.11.

4.1.5 Summary

The reported statistics should give a first global look on what linguistic information the discourse annotation physically brings. We regard this overview as a springboard for studies of various character. To name just a few, from the linguistic viewpoint, we offer the first (quantified) corpus evidence for use of Czech connectives (restricted, of course, by the given domain). In a certain extent, the annotation figures are comparable to the findings made by the PDTB team for English, and possibly to observations made for other languages with existing discourse annotation. From the computational viewpoint, the connective ambiguity analysis represents a groundwork material for automatic classification of connectives and training data for other discourse-oriented NLP tasks.

4.2 Syntax – Discourse Mismatches

One of the basic theoretical decisions for discourse annotation in Prague, as described earlier in Section 3.1.2, is the assumption of parallelism of certain syntactic and discourse phenomena and, subsequently, using advantage of relevant syntactic features in discourse processing. Which syntactic properties are relevant for discourse analysis and how are they represented in the Prague Dependency Treebank, was one of the main objectives of our previous research (Mladová 2008a) and we also mentioned them in this thesis (Section 3.1.2). Also, an in-depth study has been carried out on the nature of some basic semantic relations applying both in sentence and discourse analysis (Jínová et. al 2014).

The large-scale manual annotations of discourse relations revealed also cases, where sentence and discourse analyses differ. Such observations of differences on the syntax – discourse interface are worth addressing, as they substantially contribute to our understanding of discourse structure and coherence. The main objective of this section is to systematically describe mismatches between syntactic and discourse dependencies, as they have been documented in the data of the PDT.

The devices that enabled us to observe syntax – discourse mismatches are threefold. First, it is the inter-annotator disagreement measurement and its analysis, mainly in the initial part of the project. Second, it is the necessity to introduce the "group" value of the *range* attribute for marking untypical extents of argument spans, i.e. those that are not an exact match with dependency (sub)trees. And, third, it is the PML-TQ search engine (Štěpánek and Pajas 2010), which was in this respect used for finding untypical placement of discourse connectives.

Many syntax – discourse mismatches were identified due to the annotation inconsistencies, the most apparent one being the determining the extent of discourse arguments. A high number of such disagreements was detected through analyzing the inner-annotator agreement, as described in Zikánová et al. (2010). Most of these cases were an issue of inclusion (or exclusion) of one level in the tree, that means that one annotator believed the argument to include a governing clause and its dependent clause, whereas the other annotator marked only the dependent clause (cf. Figure 4.1: each discourse arrow was annotated by a different annotator for the same relation). Then, when we introduced a new, more liberal agreement measure that considered skipping one level in the tree structure to be still an agreement, the IAA numbers for argument extent have increased by 10%. The error analysis then showed that the governing verbs causing problems were mainly verbs of speaking, thinking, expressing attitude or verbs with very

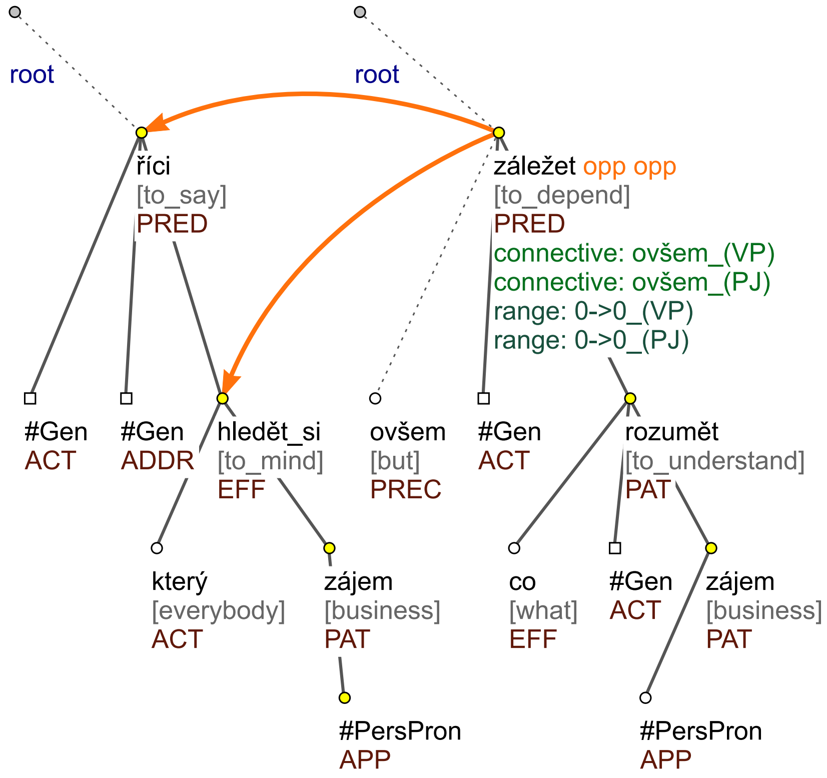


Figure 4.1: Disagreement in annotating argument extents

general or modal meanings. In this way, and in agreement with the PDTB research group, we found out that these verbs cause “a lack of congruence between arguments at the syntactic and the discourse levels” (Dinesh et al. 2005, p. 29) and can be mostly resolved with the analysis of **attribution** – the ascription of the discourse arguments and relations to the agents (sources) who expressed them. The IAA analysis of PDiT thus fully supports the findings of the PDTB group that the discourse-level annotation reveals attribution as one of the major sources of conflicts between syntactic and semantic dependencies (Dinesh et al. 2005).

Second, a detailed analysis of **discourse arguments** marked **with the “group” value** (of the attribute “start_range” or “target_range”) revealed cases where the position and extent of discourse arguments deviate from the syntactic structure. We describe and classify these non-tree-like arguments linguistically. Attribution is one of the documented cases.

Third, we observed the cases of untypical placement of discourse connectives and in particular those, where the **connective** is placed **at a distance from both of its arguments**. Such a placement of connectives is also mostly related with the notion of attribution.

The following three subsections are devoted to each of these three phenomena, respectively. Attribution in general is addressed in Section 4.2.1, other argument extent mismatches in 4.2.2 and distant placement of connectives, and in particular the phenomenon of connective raising, in Section 4.2.3.

4.2.1 Attribution

In journalism, attribution is the identification of the source of reported information. As a linguistic concept, it is mainly known (apart from authorship attribution in forensic linguistics) from sentiment analysis where identifying the opinion holder (source) is one of the main tasks (cf. e.g. Kim and Hovy 2006). In the Czech linguistic context, the concept of attribution best intersects with what is called *reprodukce prvotních výpovědí* (Karlík et al. 2002, pp. 375–376)[11], which is mainly concerned with syntactic forms of reproducing/attribution in Czech and with semantic properties of the reproducing/attribution verbs or adverbs.

In the PDTB approach, attribution is defined as a relation of "ownership" between abstract objects (Asher 1993) and individuals (PDTB manual, Prasad et al. 2007, p. 40), in other words, it is a relation of a proposition to an entity (to the person who expressed it). Thus, it is not a relation between two abstract objects (events, states etc.). That is why, according to the PDTB framework, attribution does not belong to the discourse relations sense taxonomy, as it does in the RST framework (RST-Treebank manual – Carlson and Marcu 2001, p. 10) – discourse relations (associated with an explicit or implicit connective) hold between two abstract objects.

In the PDTB 2.0, the attributed content can be a whole discourse relation (annotated for all relations indicated by explicit and implicit connective tokens or by an alternative lexicalization of the connective (altlex)) or one of the arguments only. Compare an example from the PDTB (89) where the whole *while*-relation and also its Arg1 are attributed to the writer, whereas the Arg2 is attributed to another individual, to the purchasing agents. The clause with the attribution verb (i. e. the text span complex signaling attribution) is henceforth referred to as the attribution clause.

(89) *Factory orders and construction outlays were largely flat in December* while purchasing agents said **manufacturing shrank further in October.**

[11] primary statements reproduction

Non-clausal attribution phrases of the type *podle jeho slov* (*according to*) or *prý* (*allegedly*) are not dealt with in this research even though their semantic validity is equal to the clausal (verb-containing) attribution phrases. They both do not belong to the discourse relation connecting two contents of saying and are to be excluded from the arguments. The manual annotations of both PDTB and PDiT, however, so far did not exclude these phrases from the arguments explicitly (Prasad et al. 2007, p. 15) but it is only a temporary solution. The problem arising in the annotation is demonstrated by Example (90) where the main clause of the Arg1 contains the attribution phrase *podle Kaliny* (*according to Kalina*), which, according to current annotation guidelines in both corpora, is still marked as a part of the Arg 1.

(90) Speciální kategorií je cena Komerční banky nejlepší české nahrávce. Podle Kaliny *jde o pojistku pro případ, že v ostatních kategoriích by domácí produkce neuspěla.* „**To se však nestalo,**" podotkl Kalina.

A special category is the Komerční banka prize to the best Czech recording. According to Kalina, *this is insurance for the case that the domestic production fails in all other categories.* **"However, that did not happen,"** Kalina said.

Further, four key properties of attribution were annotated in the PDTB 2.0: *source* of the attribution (the writer, other person, arbitrary source), its *type* (assertions, beliefs, facts, eventualities), *scopal polarity* (for cases with lowered negation scope, cf. Examples (91) and (92) below) and *determinacy* (for contexts where the attribution itself is cancelled by negation, conditional or infinitival constructions, cf. Example (95) below). For more details, cf. Prasad et al. 2007, p. 41.

As a contrast to the PDTB approach (treatment of attribution separately as a specific phenomenon), we present here an example of its treatment within the discourse relation sense taxonomy, following the RST and SDRT frameworks. In the German TÜBA-D/Z (Versley and Gastel 2013), a fifth major class of discourse relations is introduced, the REPORTING class. It contains two relations, which emphasize either the speaking act (Attribution) or the content (Source) in the sense of the RST notion of nuclearity (cf. Section 2.3.2.1).[12]

[12] In Source, the speaking act could be removed and the content would still make sense: *"In three million years, the earth will be rotating 1% slower", Prof. Soandso of Caltech said. = In three million years, the earth will be rotating 1% slower, as we all know.* In Attribution, the content can be replaced by an evaluation of it, and it would still make sense: *Peter claims that the moon is made of cheese. = Peter makes insane claims about the moon.*

In the PDT, the annotation of attribution is still a future task, even though thanks to some properties of the tectogrammatical analysis and the valency lexicon attached to the treebank, some attribution features are already to be found in the Czech data, like roots of a direct speech (t-layer attribute *dsp_root*), dependent content clauses of verbs of saying (marked predominantly with the functor EFF), parentheses, or attitude markers.[13]

4.2.1.1 Negation on the Attribution Verb, Determinacy

If an attribution verb in the matrix clause is negated, the negation can be interpreted "lower", within the dependent clause. This phenomenon is often referred to as negation transfer (cf. Quirk et al. 2004, pp. 1033–1034; Daneš et al. 1987, p. 266), cf. Example (91).[14]

(91) *John does not think that Mary will come.*
=> John thinks that Mary will not come.
(According to John, Mary will not come.)

The negation transfer concerns, as far as we can say, obligatorily both English and Czech basic verbs of opinion and expectation (e.g. *nemyslím si – I don't think*; *nevěřím – I don't believe*), perception (*nezdá se – it doesn't seem*) and the some verbs of saying if connected with a conditional (*neřekl bych – I wouldn't say*). From the perspective of communicative functions, negating an attribution verb in the matrix clause weakens the negation in the content clause and thus it is a form of hedging, used quite often in diplomatic and political statements, cf. the analysis of (4) from the PDT, where a Czech gymnast expresses her opinion on the performance of another, world-famous gymnast. Apparently, the negation transfer here has the function to soften the negative evaluation expressed – the original sentence (92) can be transformed to (92a) with only a little less politeness in the resulting statement.

[13] To find out, in what extent the tectogrammatical analysis contributes to attribution analysis and, thus, in what extent these features allow for its automatic treatment, is one of the directions of our next research.

[14] This mismatch between the surface structure and semantic interpretation of the negated constructions is referred to both as *Neg-raising* and *Neg-lowering* in literature, depending on whether the analyst takes the syntactic form or the semantics to be the starting point. Neg-raising would then be proceeding from the underlying semantics to a surface realization, whereas Neg-lowering would describe a process of semantic interpretation of the surface form. The English grammar book by R. Quirk et al. (2004) refers to the phenomenon simply as *transferred negation*, avoiding so the split nature of the terminology in different approaches. For more detail on the negation transfer cf. Fillmore (1963, p. 220); Lasnik (1975); Tovena (2000).

(92) *Viděla jsem Sčerba, ale nemyslím si, že on by měl být nejlepším gymnastou světa.*

I saw Scherbo, but I don't think he should be the best gymnast of the world.

(92a) *Viděla jsem Sčerba, ale myslím si, že on by neměl být nejlepším gymnastou světa.*

I saw Scherbo, but I think he shouldn't be the best gymnast of the world.

An interpretation of the negation directly in the matrix clause with the verbs (verb forms) listed above is rare, it would negate the attribution verb itself, if it stands in focus, cf. (93); verbs in focus are in capital letters:

(93) *He doesn't THINK Mary will come, he KNOWS it.*

If a discourse argument or a whole relation have a negated attribution, which reverses its polarity, in the PDTB 2.0, the feature *scopal polarity* is annotated with the value „Neg“, otherwise the default value is „Null“ (Prasad et al. 2007, p. 41).

Other verbs of attribution than those just mentioned do NOT allow for negation transfer.

(94) *John did not say that Mary will come.*

Sentences like (94), with a negation on a simple verb of saying, typically have several possible readings depending on the context. But in principle, the attribution does not hold here, the reading *John said that not A.* is not possible.[15]
Yet, even if the scope of the negation here is wide, and so the content clause itself is not negated, its validity is unsure. The attribution clause itself does not tell us anything about the validity of the content clause. In Examples (95) to (98) from PDTB and PDiT, the validity of the content is also relativized by the attribution clause (which is hightlighted in boldface), but by a different means than negation:

(95) ***It might be feared*** *that even thinking about lower budgets will hurt national security because the door will be opened to opportunistic budget cutting by an irresponsible Congress.*

[15] possible readings, and there may be more (reading number 3 is not acceptable):
1a, John did not express himself to A at all.
1b, John did not express himself to A, although everybody knows about A, he withheld A.
1c, John cannot verify A.
2, John did not say, did YELL A at somebody.
3, *John said that not A.

(96) ***It is too early to say*** *how the joint venture will be structured.*

(97) ***Opinion is mixed on*** *how much of a boost the overall stock market would get if dividend growth continues at double-digit levels.*

(98) ***Bylo by tedy možné konstatovat****, že restituce jsou brzdou privatizace.*

It would be therefore possible to state *that restitutions are a disincentive of privatization.*

In the PDTB 2.0, these indeterminate cases are marked with the „Indet“ value of the *determinacy* attribute.

Another situation arises if the attribution verb contains a negative polarity feature which, again, reverses the polarity of the content clause, as was the case with verbs (and verb forms) relevant for negation transfer. Here, the feature of *scopal polarity* is again useful to capture the narrow scope of the lexical negation in the main clause.[16]

(99) *John denied that Mary will come.*
=> John said that Mary will not come.
(According to John, Mary will not come.)

In a slightly different approach, the verbs like *deny* and *refuse* (together with verbs like e.g. *admit, concede, cause, imply, add, specify*, etc.) can be themselves viewed as a sort of discourse operators, as a specific type of alternative lexicalizations of discourse connectives. Although the notion of verbs as discourse connectives is quite unusual, they do have semantic properties that enable them to structure discourse in the same way like connectives do. For a detailed study on Czech verbs of saying in connective functions cf. M. Rysová (2014).

As already mentioned at the beginning of this section, attribution clause should not be regarded as a part of a discourse argument. The phenomena of negation transfer and indeterminacy on one hand justify a need to include an attribution clause into the argument together with the content clause since the surface negation on the attribution verb or the indeterminate nature of the attribution clause are important for the interpretation of the relation. However, the basic principle of distinguishing attribution from the attributed content makes us separate the attribution clause from the attributed content, so we consider it preferable not to include these attribution clauses in the argument, but to mark the attribution of the argument (or of the relation) with the *Neg*-value of the

[16] And again, if we negate the verb *deny*, we will end up with a wide scope of the negation on the main verb: *John did not deny that Mary will come. => According to John, it is uncertain whether Mary will come or not.* Here, we again need the „Indet“ value of the *determinacy* feature rather than the „Neg“ value of the *scopal polarity* feature.

polarity scope feature or *Indet*-value of the *determinacy* feature, as it is the case in the PDTB. In the PDiT and the PDT 3.0, no such features have been annotated so far, but it is a priority to do so for the next releases.

4.2.2 Do Discourse Arguments Match Dependency (Sub)Trees?

For the purposes of annotating text spans that do not match tectogrammatical dependency trees (or subtrees) as discourse arguments, the value “group” (of the attributes *start_range* and *target_range* of the discourse arrow) was introduced in the manual annotation (for data notation cf Section 3.2.2). This value allows for assembling any set of nodes from any tree into a group, according to annotator’s judgement about what is an argument’s position.

In this section we characterize these “non-tree-like” discourse arguments annotated in the PDiT 1.0. We first analyze them with respect to their extent and location and then according to their syntactic and semantic properties. We also look at the nature of the material left out of the argument.[17]

The value “group” was NOT used to gather a certain number of full sentences into an argument. For full sentences, the values of *start/target_range* attribute are numbers that determine the argument extent, “0” meaning that the argument only includes the actual sentence (subtree), “1” meaning the arguments includes the actual sentence (subtree) and one full sentence following in the linear text order, and so on. This implies that the device of grouping only treated arguments where some text from at least one of its sentences was left out or the argument consisted of non-adjacent sentences.

There are 262 instances of discourse relations with one or two group-like arguments in the PDiT 1.0 data.[18] For the analysis of groups, we manually checked a random 100 of them in order to describe the basic tendencies.

One remark at the beginning should be made, though, which concerns the technical aspects of group annotation. Initially, if a discourse relation existed between a governing and a dependent clause within one sentence, it was clear that the governing clause argument does not include the dependent clause argument. This is not quite straightforward, as the tectogrammatical dependency structures are understood to project all the way down, and including any parent node in

[17] We are aware of the fact that we do not speak here about all surface discontinuities of discourse arguments. Some surface discontinuities can be still covered by tree-like structures, albeit non-projective ones.

[18] In this section, the research was conducted on the 9/10 of the treebank in order to leave the testing part of the dataset unobserved.

a discourse argument means also including all its children. Therefore, in these cases, a group was created for the parent (governing clause) in order to avoid the children nodes (dependent clause) from building an intersection of the two arguments. Later, in course of the annotations, creating this type of groups was canceled due to its complexity which was slowing down the annotators. It was decided that even without group marking, two discourse arguments in the relation of syntactic dependency will be understood not to contain each other. The number of relations with grouped arguments is therefore lower in the updated PDT 3.0 (177 instances).

Regarding the location of the "grouped" arguments, from the observed sample of discourse groups, the following tendencies are apparent. (We divide our observations to groups annotated within a single sentence (A) and across sentence boundaries (B) and demonstrate every case with a treebank example. The grouped arguments are highlighted in blue.

A. Grouping of the nodes used **within a single sentence** mainly covers the following five types of structures. The last two cases mentioned below are of technical rather than linguistic nature: they demonstrate a way of solving a specific issue of the tectogrammatical representation (ellipsis restoration, multi-element coordination).

1. Relation of a parenthesis to a part of the sentence (the non-related part of the sentence must be excluded from the argument. The annotation of the parenthesis in Example (100), mentioned already earlier in Section 3.2.2 under (22), is represented by Figure (3.2) in the same Section.

(100) *K pěstování vědy je třeba nejen střecha nad hlavou, nějaké finance* (a **někdy jich je třeba dost**), ale především vědecký dorost.

For cultivation of science, it is necessary to have not only a roof over your head, some finances (and **sometimes there needs to be plenty**), but especially young researchers.

2. Leaving out a member of a coordination, which is not relevant for the discourse relation coming next:

(101) *Zánik takové smlouvy je možný* buď dohodou obou stran, anebo výpovědí jedné strany, případně *odstoupením od smlouvy,* pokud **je tento způsob ukončení smluvního vztahu ve smlouvě sjednán.**

Termination of such a contract is possible either by agreement of both parties, by dismissal of one party or *by withdrawal from the contract* if **this**

way of terminating the contractual relation has been negotiated in the contract.

3. Leaving out a dependent clause. Examples and further description cf. below under (107) to (109).
4. Syntactically unanchored direct speech: In some cases, direct speech is not syntactically integrated in the sentence, since the valency frame of a possibly introductory verb is already full (compare the verb *vysvětlovat – to explain* in Example (102)). In such cases, in the tectogrammatical analysis a complement node with an empty lemma (#EmpVerb) is generated, which is then the governing node for the direct speech. The meaning of the node is roughly *řka* (*saying that*). The device of grouping in discourse annotation then excludes this node from the discourse arguments. The reconstructed node builds another verbal unit, which is then related to the dependent clause via attribution, not via discourse relation. The reason of the grouping here is therefore rather a technical matter.

(102) *Nepřesvědčivý výkon vysvětloval trenér Vladimír Vůjtek*: **“Protože jsme postup do play off vybojovali už dříve, hráli jsme dvě třetiny na čtyři útoky.**

The unconvincing performance was explained by the coach Vladimir Vůjtek: “**As we have won the advancing to the playoffs already earlier, we played two thirds of the match with four different formations.**

5. Gathering two or more coordinated clauses which are in a discourse relation to another clause within the same multiple coordination. This is another technical matter of the tectogrammatical tree representation, which allows for coordination of more than two members.

(103) **Přišlo málo lidí, my jsme přesto hráli** a pak *když nám měl zaplatit, chyběly mu tři tisíce.*

Not many people came, we played nevertheless, and then *when he had to pay to us, he lacked three thousand crowns.*

B. Grouping of the nodes is used **for more sentences than one** mainly concerns the following two structures:

1. Discourse argument is larger than a single sentence; it can contain any number of full sentences, but at least one sentence only partially. Most common is a discourse argument containing one full sentence and a part of an adjacent sentence. This case is demonstrated by Example (104) and Figure 4.2.

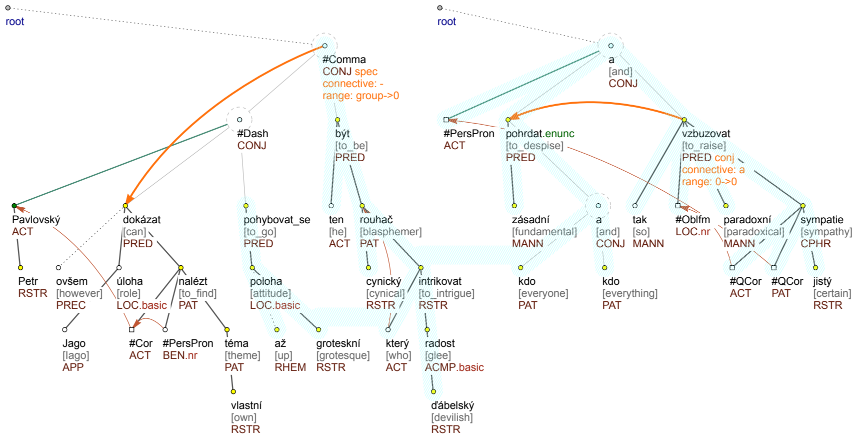


Figure 4.2: Group annotation in more sentences

(104) *Petr Pavlovský si ovšem v úloze Jaga dokázal nalézt vlastní téma* – **pohybuje se v až groteskních polohách, je to cynický rouhač, který intrikuje s ďábelskou radostí. Zásadně pohrdá všemi a vším a paradoxně tak vzbuzuje jisté sympatie.**

Petr Pavlovský, however, could find his own theme in the role of Iago – **he goes up to grotesque attitudes; he is a cynical blasphemer who intrigues with devilish glee. He fundamentally despises everyone and everything, and so paradoxically raises certain sympathy.**

2. The argument contains several sentences or parts of them, which are non-adjacent:

(105) *Šedesát tři vězňů, kteří vykonávají trest odnětí svobody v České republice, požádalo za první půlrok o předání do věznic na území Slovenska.* Informoval o tom včera tiskový mluvčí generálního ředitelství Vězeňské služby ČR Eduard Vacek. Dodal, *že loni podalo tuto žádost 200 odsouzených.* **Praktické předávání** však **začalo až letos v červnu, kdy bylo předáno 16 odsouzených.**

During the first six months, sixty-three prisoners who serve their sentence in the Czech Republic asked for handing over to prisons in Slovakia. The spokesman of the Directorate General of the Prison Service of Czech Republic Eduard Vacek announced this yesterday. He added *that last year, 200 inmates filed the request.* But **the transfer itself started only in June this year, when 16 inmates were handed over.**

So far, we have described the non-tree-like arguments with respect to their extent and location. Now we will provide syntactic and semantic characteristics of such grouped arguments, or, more precisely, name the reasons why these arguments do not match a dependency tree, mostly by characterizing the material excluded from the argument. The part of the sentence which is typically excluded can be divided into three categories. The remaining grouped arguments were isolated cases with rare structures.

1. First and most commonly, it is clauses of attribution – that means statements with verbs of saying and thinking introducing a reported content, cf. Example (106). For this type of arguments (with an excluded attribution clause) it may happen that the discourse connective occurs outside of both the arguments, namely inside the attribution clause. We call this phenomenon *connective raising* and describe it in more detail in Section 4.2.3.1 of this thesis.

2. It is various types of parentheses, typically structures of the type *as they say, as we know, as we can see* or *as was said/mentioned/*etc., cf. Example (107). These structures mostly also belong semantically to attribution statements but their syntactic structure is fixed and, unlike introductory statements, they do not govern the whole sentence.

3. Finally, certain dependent clauses can be excluded from the argument. Annotators treated them in these cases as not necessary for establishing the semantics of the respective discourse relations. We found examples of excluded concessive (107), descriptive attributive (108) and contrastive (109) dependent clauses.

(106) *Britská vláda učinila včera další krok ke zmírnění napětí v provincii,* když ministr pro Severní Irsko oficiálně oznámil, že **v ulicích severoirských měst byla zredukována přítomnost britských vojáků a policejních sil.**

The British government made a further step yesterday to ease tensions in the province, when the minister for Northern Ireland officially announced that **the presence of British troops and police forces in the streets of Northern Ireland's cities had been reduced.**

(107) Přestože klusácký sport u nás dosahuje při srovnání zahraničních startů daleko větší úspěšnost než cvalový provoz, *stál pro téměř neexistující propagaci na okraji zájmů dostihové veřejnosti.* Jak však včera zaznělo na tiskové konferenci, **měla by se tato situace letos změnit k lepšímu, protože ČKA rovněž podepsala smlouvu o marketingu se společností Impact (dceřiná společnost Art production K.).**

Although harness racing in our country achieves far greater success than gallop racing when comparing foreign starts, *because of almost nonexistent promotion it was marginal for the racing public.* However, as was said yesterday at a press conference, **the situation should change for the better this year, since the ČKA also signed a contract about marketing with the Impact company (a subsidiary of the Art Production K.).**

(108) (Context: K tragédii na okruhu formule 1 se nevyjádřila stáj zesnulého jezdce A. Senny.)

To ostatně neudělal ani čtyřnásobný mistr světa Alain Prost, kterého po skončení kariéry Senna loni u týmu Williams Renault nahradil, zato **opět vystoupil s kritikou odpovědných za apatii vůči bezpečnosti závodníků.**

(Context: The racing stable of the deceased F1 rider A. Senna did not comment on the tragedy on the circuit.)

Neither did the four-time world champion Alain Prost, who was replaced at the end of his career last year by Senna in Williams Renault team, yet **he came out again with criticism of people responsible for the apathy towards the safety of the racers.**

(109) Zatímco u dospělých dojde při kašli k pouhému stáhnutí svalů, **dětské dýchací cesty se zalijí hlenem a vypadá to, že dítě má astmatický záchvat.** *Když* pak *povyroste, ukáže se, že to nemusí být astma bronchiale, a příznaky zmizí,* říká Tamara Svobodová.

While in case of adults only muscle contractions occur during coughing, **a child's respiratory tract fills with mucus and seemingly the child has an asthma attack.** Later*, when it gets bigger, it turns out that it may not be a bronchial asthma, and the symptoms disappear,* Tamara Svobodová says.

The device of grouping in the PDiT 1.0 and the PDT 3.0 annotation demonstrates the same tendency in the data that we have already observed earlier in this section: while different structures can take part in creating “non-tree-like” discourse arguments, it is again predominantly (various realizations of) attribution, that is responsible for the major mismatches between syntactic and discourse dependencies.

4.2.3 Distant Position of the Connective from its Arguments

Typical placement of discourse connectives in Czech is well known: subordinating conjunctions are to be found within the clause they introduce, e.g. within one of the discourse arguments they relate; coordinating conjunctions typically take position between their two arguments, the arguments being adjacent. Discourse adverbs, particles and PPs connecting inter-sentential relations appear mainly at the beginning of or within the second argument in the surface order, and their first argument (in the surface order) can be adjacent or distant, as shown by Webber et al.(2003).[19]

Looking at positions of connectives deviating from these regular patterns is another possibility for finding source of conflicts between the syntactic and the discourse analyses. In the PDiT 1.0[20], we detected 72 cases of discourse relations where the connective was placed at a distance from both of its arguments.[21]

Such a case can be demonstrated by Example (110) where the connective *ale* (*but*) is placed in the main clause of the last sentence. The main clause itself is not a part of any of the arguments here.

(110) Všichni, kdo jsou spjati s divadly, vědí, že *vícezdrojové financování je jedině možné a správné.* Vědí ale také, že **pro to v naší společnosti dosud nejsou vytvořeny podmínky.**

[19] Webber et al. (2003) analyze English discourse adverbials. Nevertheless, in this respect, we believe that the general syntactic principles of conjunction placement and principles described by Webber et al. for discourse connective placement could be regarded the same for Czech and English (and also for some other languages.)

[20] again, 9/10 of the data

[21] The best approximation to finding the revelant connectives was achieved with the following pml-tq query.

```
t-node $t :=
  [ !descendant $n4, !sibling $n3, !parent $n4,
    member discourse
    [ t-connectors.rf t-node $n4 :=
      [ functor != "RHEM",
        0x parent t-node
          [ nodetype = "coap" ] ],
      target_node.rf t-node $n3 :=
        [ !descendant $n4, !parent $n4 ] ] ];
```

Still, some results had to be sorted out as irrelevant and a few were found with help of additional queries. This concerns structures introduced with *jak* (*as*). Also, the borderline of this analysis is set by some inconsistency in the annotation of distant connective placement, caused by the complexity of the phenomenon.

Type of distant placement		**% (100% = 66)**	**two relevant interpretations (%)**
clauses with verbs of attribution	connective raising – verbs of saying and thinking	27.3	12.1
	connective raising – verbs with general meanings, of existence etc.	16.7	16.7
	As-constructions: reversed syntactic order	1.5	
DC as a part of an altlex construction		1.5	
DC in questions and answers		3.0	
DC in a separate sentence		4.6	
irrelevant material		16.7	

Table 4.12: Distant placement of discourse connectives in PDiT 1.0

> All those closely related to theaters know that *multi-source funding is the only possible and correct way.* But they also know that **in our society, there have not yet been established conditions for this.**

A detailed syntactico-semantic analysis of the "distant connectives" follows; the quantitative results for PDiT 1.0 are summed up below in Table 4.12. From the 72 results of the query mentioned above, 6 were found twice (multiword connectives) and further 9 showed to be irrelevant for our study (list structures, wrong annotation or awkward style).

The core of the "distant connectives" can be then divided into four groups according to their structure:

1. The distant position of a connective brings us back to the notion of attribution: The highest percentage (74.3%) of the distantly placed connectives discovered was documented in clauses with attribution verbs. Typically in these constructions, the connective syntactically belongs to the attribution clause but semantically it belongs to the lower, dependent clause (content clause), as in Example (110) above. The attribution clause is therefore not a part of any of the arguments. We have witnessed this phenomenon already earlier in the analysis of "non-tree-like" arguments; we call it connective raising (henceforth also CR) and address it in detail further in Section 4.2.3.1.

Verbs of attribution, in the sense the PDTB 2.0 approach classifies them, are verbs of saying and thinking expressing assertions, beliefs, facts or eventualities

(Prasad et al. 2007, p. 44). Nevertheless, the same type of distant placement of a connective was in the PDiT 1.0 documented also for structures expressing speaker's attitude towards the content in an indirect, general or impersonal way, such as: *je zřejmé* (*it is clear*); *zdá se* (*it seems*); *předpokládá se* (*it is assumed*); *je možné* (*it is possible*); *dá se říci* (*we can say*); *je třeba si uvědomit* (*it should be realized*); *není se co divit* (*no wonder*) etc., cf. Example (111).

Therefore, within this group, we further differentiate between CR with typical attribution verbs (of saying and thinking) and verbs with general meanings of existence, uncertainty or modal modifications. Distinguishing between those two groups proved convenient, as the certainty about what happens semantically in those two types of structures is very different. We show this in Table 4.12: numbers in the rightmost column indicate the percentage of cases for every type of structure where it was difficult to decide between a CR and non-CR interpretation, both were relevant. This issue is described in more detail in Section 4.2.3.1 on connective raising.

(111) *Díky mnohým pokoutním překupníkům bylo totiž v uplynulém období hodně cizinců napáleno* a tak se není co divit, že **teď jsou všichni potenciální kupci hrozně opatrní.**[22]

Because of many traffickers many foreigners were duped lately, and so it is no wonder that **now all potential buyers are very cautious.**

Regarding the syntactic properties in this group, we found two types of constructions where attribution clauses are to be excluded from the argument extent. The first, a most common structure, is exemplified above as (110) – the attribution clause is the governing clause of the second argument in surface order. The second structure is an inversed syntactic relation between clauses, compare Example (112);[23] in Czech it is typically introduced with a semantically underspecified expression *jak* (*as*): *Jak řekl...* (*As he said...*).

(112) „V lize se objevím až na podzim, *tak mám alespoň čas věnovat se pořádně mé přítelkyni.*“ Jak ale Šmicer dodal**, nucenou pauzu nevyužije jako jeho spoluhráč Berger - na ženění je prý brzy...**

"I won't show up in the league until the autumn, *so at least I have time for my girlfriend.*" But as Šmicer added, [lit: As but Šmicer added,] **he won't**

[22] In our opinion, according to the Czech orthography rules, there should be a comma before the *a* here indicating a different relation then a simple conjunction.

[23] in the PDT, these constructions are treated as a special type of parenthesis (cf. Mikulová et al. 2005, p. 305).

use the forced break like his teammate Berger - is it too soon get married...

The remaining three groups are quite small, and they also reinterpret the notion of a „distantly placed" connective in a certain way. They should be viewed as marginal.

2. In group 2, the connective is a part of an alternative lexicalization phrase (”altlex", Section 2.3.2.2), cf. Example (113). This means that an altlex phrase, which could be represented even by a separate verbal clause, is never an argument itself. In this way, the connective itself is distant from both its arguments, even though the altlex expression can be adjacent to both its arguments.

(113) *U našeho auta poněkud zlobilo nepříliš poddajné řazení,* ale to lze přičíst skutečnosti**, že jsme vyjeli s vozem, který neměl najeto ani celých čtyřicet kilometrů.**

((to) lze přičíst skutečnosti = (to) proto)

Our car annoyed us a little with the not very pliable gear shifting, but this can be attributed to the fact **that we drove out with a car that had barely forty kilometers on the clock.**

((this) can be attributed to the fact that = (this is) because)

3. A special group is represented by connectives in question – answer pair, cf. Example (114).

(114) *Naši dnešní policii zajímá leccos.*
Ale občan v tísni, zvláště ten nebohatý, bez vlivu, bezbranný?
Ten ji zajímá jen málo.

The police of today are interested in many things.
But a citizen in need, especially the poor, with no influence, defenseless?
That one concerns them little.

From the nature of question – answer relation it is clear that we deal mostly with elliptical structures. Example (114) was interpreted by the annotator as a relation between two declarative sentences, whereas the connective (*ale – but*) is placed in the question in between, which does not belong to any of the arguments. The contrastive relation between the police being interested in many things and their paying too little attention to the actual problems of a citizen in need is interrupted by asking about the citizen in between. The choice of the annotator to exclude the interrogative sentence from the second argument is one possible solution.

The other solution here is to pay attention to the coreference link of the pronoun *ten* (*that one*) in the second argument. It brings us back to the interrogative, and so to say merges the question with the answer into a single semantic unit, which is then more useful viewpoint: "But the police of today is little concerned with the citizens in need." In that case, the question would be included in the second argument and the connective would not be distantly placed any longer.

4. In a way similar to the third group, connectives appearing as the only element/constituent of a sentence, cf. Example (115), are cases that are "distant" from their arguments in the way that there is a sentence boundary between them and the arguments they connect. This can be viewed as a special type of parceling. Such a process of utterance segmentation is more usual for sentence constituents in the focus part of the sentence, but, also here for the connective, it has the primary function to underline and emphasize the detached element.

(115) Pláč muzeí. *Zaslechnete ho vlastně v každém takovém nevýdělečném zařízení - nejsou peníze.* A přece... **Na příkladu Technického, ale i Národního muzea Profit na své třetí straně ukazuje, že vydělat tyto nevýdělečné organizace mohou.**

Weeping of the museums. You can hear it in any such non-profit facility – there is no money. And yet... **On its page 3, the Profit magazine shows on the example of the Technical as well as the National Museums that these non-profit organizations can in fact earn.**

4.2.3.1 Connective Raising

As already stated, the first, largest group, represented by the first line in Table 4.12, contains structures with attribution and external connective placement:

In Example (110) above, the attribution clause in the second sentence *Vědí ale také* (*But they also know*) contains two connectives, one of which (*také – also*) relates the two matrix clauses with a conjunction relation, whereas the other, the connective *ale* (*but*) can only be interpreted "lower" in the structure, as a contrast between the two subordinate (content) clauses, not between the two occurrences of the verb *vědět* (*to know*).

Main clauses plan:

They know A. – They also know B.
**They know A. – But they know B.*

Subordinate clauses plan:

Multi-source funding is good. But there are no conditions for it yet.

This phenomenon is henceforth called **connective raising:**[24] A (usually con-

[24] We are fully aware that the phenomenon of negation transfer described in Section 4.2.1.1 above is analogical to the one described here – a surface form from a matrix clause is in

trastive) connective is placed outside of both of its arguments, it syntactically belongs to the matrix clause but, from the semantic viewpoint, it is interpreted in the dependent clause.

Wide and narrow connective "scopes"

Structures with connective raising are not to be confused with other cases where the connective in fact syntactically and semantically relates two clauses of attribution (including their respective content clauses). This was demonstrated by Dinesh et al. (2005) with the English example from PDTB (116) and is also visible on the Czech example (117) from the PDT.

(116) *Advocates said the 90-cent-an-hour rise, to $4.25 an hour by April 1991, is too small for the working poor*, while **opponents argued that the increase will still hurt small business and cost many thousands of jobs.**

(117) *Prezident Havel při odchodu z jednání vlády uvedl, že v diskusi zvítězil názor, aby na oslavy byli přizváni i zástupci Německa.* **Premiér Václav Klaus** však **po skončení jednání LN řekl, že vláda o této věci včera ještě nerozhodla.**

When leaving the government meeting President Havel stated that the view prevailed in the debate that the representatives from Germany should also be invited to the celebrations. **Prime Minister Václav Klaus,** however**, said to the LN after the meeting that yesterday the government had not yet decided on the matter.**

Determining whether the connective actually semantically relates the attribution clauses or whether it operates at a lower level between the attributed contents, is a problematic issue. The PDiT annotators were asked to use an intuitive transformation test which proved to be quite helpful, yet not always univocally decisive:

The "scope" of the connectives in attribution clauses can be tested with: (i) moving the connective lower, to the content clause and (ii) leaving out the attribution clause in the second argument (in the surface order) while preserving the connective in question. If the connection of content clauses related with an (originally distant) connective preserves the original meaning, we can speak about connective raising. On the other hand, if such a connection gets semantically weird or incomplete, the connective relates the governing, attribution clauses.

both cases semantically interpreted in the dependent clause. Nevertheless, we chose to use the term *connective raising* to indicate that, proceeding from the discourse analysis point of view, we take the semantic level to be the starting point of our considerations.

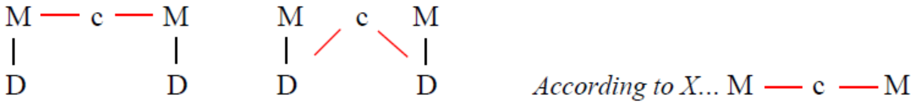


Figure 4.3: "Scopes" of connectives in sentences with attribution

The different situations are illustrated by Figure 4.3: In the left part, the connective ("c" in the figure) operates at the level of main clauses ("M"); in the middle part, it operates at the level of dependent ("D") clauses (connective raising); and in the right part, a flat structure with the non-clausal *according to X* phrase is presented. Compare the situations under (i1-ii1) and (i2-ii2), where the tests are applied for the original corpus examples mentioned above as (110) and (117), respectively.

(i1) lowering of the connective: the connective operates at the level of attributed contents

(118) Všichni, kdo jsou spjati s divadly, vědí, že *vícezdrojové financování je jedině možné a správné.* Vědí také*, že* **pro to v naší společnosti ale dosud nejsou vytvořeny podmínky.**

All those closely related to theaters know that *multi-source funding is the only possible and correct way.* They also know that **in our society, nevertheless, there have not yet been established conditions for this.**

(ii1)leaving out the attribution clause: the connective operates at the level of attributed contents

(119) Všichni, kdo jsou spjati s divadly, vědí, že *vícezdrojové financování je jedině možné a správné.* **Nejsou pro to ale v naší společnosti dosud vytvořeny podmínky.**

All those closely related to theaters know that *multi-source funding is the only possible and correct way.* **However, in our society, there have not yet been established conditions for this.**

(i2)lowering of the connective: awkward. The connective operates most likely at the level of main clauses.

(120) *?Prezident Havel při odchodu z jednání vlády uvedl, že v diskusi zvítězil názor, aby na oslavy byli přizváni i zástupci Německa.* **Premiér Václav**

Klaus po skončení jednání LN řekl, že vláda však o této věci včera ještě nerozhodla.

?*When leaving the government meeting President Havel stated that the view prevailed in the debate that the representatives from Germany should also be invited to the celebrations.* **Prime Minister Václav Klaus said to the LN after the meeting that yesterday the government, however, had not yet decided on the matter.**

The placement of the connective *však* (*however*)in (120) appears disruptive in the dependent clause. This is most likely because we expected it to come earlier in the sentence (or, for that matter, higher in the sentence structure). The contrast here apparently relates to the person of Prime Minister claiming something else than the President, and the contrastive connective tends to stand in close proximity.

(ii2) leaving out the attribution clause: the connective operates at the level of attributed contents, but with a loss of important information. This transformation is thus not possible.

(121) *Prezident Havel při odchodu z jednání vlády uvedl, že *v diskusi zvítězil názor, aby na oslavy byli přizváni i zástupci Německa.* **Vláda však o této věci včera ještě nerozhodla.**

*When leaving the government meeting President Havel stated that *the view prevailed in the debate that the representatives from Germany should also be invited to the celebrations.* **However, the government had not yet decided on the matter yesterday.**

The example (121) makes perfect sense even after the transformation but the meaning has shifted: the omitted information that the second argument is a (contradictory) statement of the Prime Minister, not the President, is an important one. In this case, the connective *však* (*however*) operates between the attribution clauses, which means, the syntactic and the discourse interpretations go hand in hand.

As far as we could observe, a contrastive connective syntactically connecting two main attribution clauses takes the wide scope, that means the syntactic and the discourse interpretations match, in structures with non-identical sources of attribution. In other words, the connective must take a wide scope, if the two contents in dependent clauses are expressed by two different agents (as in (116) and (117) above). A little more complicated is the situation in cases where our test (neither moving the connective lower nor leaving out the attribution clause) does not rule out any of the readings as awkward or incomplete.

(122) V České republice působí v současné době osm až deset větších nakladatelství, která se zabývají převážně vydáváním učebních textů. Konkurence sice prospívá kvalitě, učitelé i žáci mají zpravidla také větší možnost výběru než dřív, občas však dojde i k výpadku. *Letos například chybí důležitá učebnice českého jazyka pro šesté třídy.* Zástupci nakladatelství Fortuna ale tvrdí, **že by měla být k dispozici nejpozději v říjnu.**

In the Czech Republic, there are currently eight to ten major publishing houses engaged primarily in issuing textbooks. Competition is good for quality and teachers and pupils have usually also bigger choice than ever before but occasionally, there are supply disruptions. *This year, for example, an important Czech language textbook for the sixth grade is missing.* The representatives of the Fortuna publishing house however claim that **it should be available by October at the latest.**

As Example (122) shows, leaving out the attribution clause *Zástupci nakladatelství Fortuna tvrdí* (*The representatives of the Fortuna publishing house claim*) and moving the connective to the content clause does not violate the overall meaning in such a way as in (121) above. It seems that the importance of exactly Fortuna representatives making a positive claim about the availability of the textbooks is lower than the importance of the Prime Minister Klaus opposing the President's claim. In our view, the information itself (that the textbook will be available soon) suppresses the importance of the source claiming it. In this respect, both the interpretation with the wide scope of the *ale*-connective and the narrow scope (raised connective) are relevant here. Even though we arrive here at the borderline of the ability to determine one correct interpretation, we can imply an important fact here: connective raising of contrastive connectives takes place in cases where there is only a single source or where a second source of the contrasting claim is not important.

The analysis of attribution clauses with connectives in the PDiT data, i.e. of structures where the annotators either left out the attribution clauses from the arguments or where they included them, has also led to the following observations.

(i) The vast majority of structures with connective raising concerns relations with connectives from the CONTRAST class. However, despite the previous findings of M. Rysová (2014, p. 934) that connective raising only appears in structures with contrastive connectives, we were able to document a small number of connective raising cases with other meanings: 5 cases from the CONTINGENCY class (cf. Example (111) above), 4 cases from the EXPANSION class (Example (125) below) and one from the TEMPORAL class (123) in the PDiT.

(123) *Předepsal mu uklidňující a tlumivé léky.* Údajně to nakonec dopadlo tak, že **se stal jeho osobním ošetřovatelem a zároveň i nenápadným dohližitelem.**

He prescribed tranquilizing and sedative drugs to him. Reportedly, it eventually turned out, **he became his personal nurse and also a discreet supervisor.**

(ii) The high number of contrastive relations among the connective raising cases was detectable due to the opposite polarity (contrastive nature) of the lower verbs (and arguments) being related. The attribution verbs are typically of the same polarity and thus a contrastive connective semantically does not fit there.

(iii) Detecting connective raising in non-contrastive structures is more complicated, since the scope of the connective cannot be semantically ruled out at any of the levels. That is also why the interpretation is not always that clear as with the contrastives.

(iv) Regarding the connectives of conjunction, it seems to be of importance, whether the main clauses are indeed clauses of attribution (124), or whether the second main clause expresses some degree of modality or an attitude of the speaker/writer towards the content of the dependent clause (125). In the first case, the case of a clearly attributed two arguments, it seems undecidable by linguistic means, whether a conjunctive connective relates the attribution clauses or solely the content clauses. An implicit conjunction can be postulated here between both syntactic levels. In the second case, if the main clause expressing attitude or modality does not influence the semantic validity of the dependent clause (or, more precisely, if it has the same polarity as the dependent clause), as in Example (125) *je třeba si uvědomit*, the main clause was left out by the annotator as a vague attribution-like material not relevant for the relation, which is a fair interpretation.

(124) *Jsou to věci, které by se neměly stát v žádné bance, řekl ředitel pražské pobočky Banky Haná Jan Rolc* a **dodal, že banka musí být především důvěryhodným ústavem, který jedná s rozvahou.**

These are things that should not happen in any bank, said the director of the Prague branch of the Banka Haná Jan Rolc and **he added that a bank must especially be a credible institution that acts with deliberation.**

(125) *Uchytit se v této konkurenci není a nebude snadné, pokud to bude vůbec možné.* V neposlední řadě je si třeba uvědomit jedno: **pro všechny lati-**

noameričany je prioritou Severoamerická zóna volného obchodu (NAFTA) a připojení se k ní, nikoli izolované společenství s jednou středoevropskou zemí.

To hold up in this competition isn't and won't be easy, if possible at all. Finally, it is important to realize one thing: **for all Latin Americans, the North American Free Trade Agreement (NAFTA) and joining it is a priority, rather than an isolated partnership with one Central European country.**

(v) Similar is the situation with the few marked occurrences of connective raising with causal connectives. In the example (111) above, repeated here for convenience as (126), the attitude clause can be left out (and potentially replaced by an adverb of attitude, as in (21a)) and the original causal meaning is preserved. The five documented cases nevertheless differ in their structure and interpretation, and for a more detailed analysis more documented material would be needed. These cases have the various interpretations: illustrating speaker's /writer's attitude towards the validity of a causal relation – the attribution then relates to the whole relation (126), showing the speaker's /writer's implication from (127) or justification of (128) an event or previous assertion – the attribution only relates to the second argument in surface order. In the example (129), the writer justifies his/her view of sequence of events by somebody else's assertion.

In all these cases, a substitution of the main attribution clause with an adverb of attitude or an *according-to* phrase is possible with no shift in the meaning.

(126) *Díky mnohým pokoutním překupníkům bylo totiž v uplynulém období hodně cizinců napáleno* a tak se není co divit, že **teď jsou všichni potenciální kupci hrozně opatrní.**

Because of many traffickers many foreigners were duped lately, and so it is no wonder that **now all potential buyers are very cautious.**

(126a) *Díky mnohým pokoutním překupníkům bylo totiž v uplynulém období hodně cizinců napáleno, a tak* **jsou teď (pochopitelně) všichni potenciální kupci hrozně opatrní.**

Because of many traffickers many foreigners were duped lately, and so **now, (understandably), all potential buyers are very cautious.**

(127) Pavel Hirš (LSNS), poslanec: *Válku vyhrály vítězné mocnosti a Německo snad takovou mocností nebylo.* Proto si myslím, **že by Němci neměli být pozváni.**

Pavel Hirš (LSNS) a deputy: *The victorious allied powers won the War and Germany was not such a power.* Therefore I think **that the Germans shouldn't be invited.**

(127a) **Proto by (podle mne) Němci neměli být pozváni.**

Therefore (according to me) the Germans shouldn't be invited.

(128) **Souboj dvou skupin, které zatím nenašly společnou řeč v zákulisí (viz třenice o propojení fotbalových klubů s obchodními společnostmi během valné hromady ČMFS), se ale na trávníku neuskuteční.** *„Začínají jarní prázdniny a já budu s dětmi na dovolené v Alpách. Termín se nehodí ani dalším lidem z fotbalového svazu,"* prohlásil totiž včera nečekaně předseda ČMFS František Chvalovský.

A duel of the two groups that had not yet found a common ground offstage (see the fights about merging the soccer clubs with commercial companies during the general meeting of ČMFS), will however not take place. *"The spring holiday begins and I'll be with my kids on vacation in the Alps. The date doesn't fit to even more people from the Soccer Association,"* the association chairman Francis Chvalovský said unexpectedly yesterday [as a matter of fact].

(128a) *„Začínají totiž prázdniny..."*

"The spring holiday begins [as a matter of fact]..."

(129) *Neméně konfliktní situace však nastane, jestliže Sejm přece jen prezidentovo veto přehlasuje,* neboť Lech Walesa už prohlásil, **že ani v tomto případě nehodlá zákon akceptovat.**

But an equally conflict situation can occur if the Sejm outvotes the President's veto after all since Lech Walesa has already stated **that not even in this case he does intend to accept the law.**

(129a) **...neboť Lech Walesa dle svého prohlášení nehodlá zákon akceptovat.**

...**since Lech Walesa according to his statement does not intend to accept the law even in this case.**

Why Does a Connective Rise?

So far, we have described cases where a discourse connective is interpreted lower in the dependency structure than where it appears on the surface. In this subsection, we try to answer the question WHY this happens.

It would be only natural to assume that structures with connective raising are marked, whereas structures with typical placement of the connective, that means, inside one of the arguments or immediately adjacent to the arguments, are unmarked.

The examples (130a) to (130d) below, abbreviated and transformed versions of (110) above, are all grammatical in Czech. In the examples (a) and (d), the connective *ale* (*but, however*) is "raised" while in (b) and (c) it occurs in the dependent clause, where it also semantically belongs.[25]

(130) *Všichni divadelníci vědí, že vícezdrojové financování je jedině možné a správné.*
(a) Vě*dí ale také, že pro to dosud nejsou vytvořeny podmínky.*
(b) *Vědí také, že pro to ale dosud nejsou vytvořeny podmínky.*
(c) *Že pro to ale dosud nejsou vytvořeny podmínky, vědí také.*
(d) *Že pro to dosud nejsou vytvořeny podmínky, vědí ale také.*

All those closely related to theaters know that multi-source funding is the only possible and correct way.
(a) *However, they also know that there haven't been established conditions for this yet.*
(b) *They also know that there, however, haven't been established conditions for this yet.*
(c) *That there, however, haven't been established conditions for this yet, they also know.*
(d) *That there haven't been established conditions for this yet, they, however, also know.*

The almost 50 000 sentences of the PDT are not enough to observe the distribution of the "raised" vs. "non-raised" connectives. We used the 2.7 billion tokens – 178,499,792 sentences of the SYN corpora of the Czech National Corpus (Hnátková et al. 2014)[26] to see the basic preference in using these structures (with contrastive connectives) in written contemporary Czech.

We compared the distribution of the following two patterns:

[25] The English translations may not be the best translations, but, again, they are the best possible approximations to the original sentences.

[26] and www.korpus.cz. Further on, we also use the abbreviation CNC.

(i) [The beginning of the sentence – verb – (0-3 positions) – connective “však”/”ale” – comma – subordinating conjunction “že”][27], exemplified here with the following structures:

= **Řekl ale, že...**
But he said that...
Lit.: He-said but, that...
= **Doplnil jsem však, že...**
I have added, however, that...

(ii) [The beginning of the sentence – verb – (0-3 positions) – comma – subordinating conjunction “že” – connective “však”/ “ale”], exemplified as:

= **Řekl, že ale...**
He said that, however, ...
Lit.: He-said, that but...
= **Uvedli jsme, že však...**
We have stated that however...

The results are documented in Table 4.13. We have divided the results to structures with one to three arbitrary positions between the verb and the connective/comma (which allows for inclusion of most multi-word verb forms) – first 6 rows of Table 4.13[28], and to fixed single-verb structures – last 7 – 10 rows of Table 4.13. For both of them, it is clearly visible that structures with raised connectives are more frequent than structures with lower placement of the connective. Both absolute and relative numbers of occurrences differ significantly for structures with raised and non-raised connectives.

The results of the distribution analysis indicate that a (contrastive) connective tends to “climb” quite often. These structures are also not only regular, but even more frequent than the lower placement of the connective. After inspecting sample result texts from both corpora in detail, we can also suggest **why** connective raising happens: Quite simply, concerning the basic function of a discourse

[27] example of a KonText query for the first row of Table 4.13
<s> [tag=”V.*”][]{0,3}”ale””,””že”

[28] Rows 5 and 6 represents structures where both in the main clause and in the dependent clause one to three arbitrary positions are allowed – we added these two rows to the table in order to reflect the possible positions of Czech clitics (= in front of the connective).

Query	Pattern	Example of the structure[29]	Raised DC	Occurrences	i.p.m.[30]
1.	(i)	Řekl (...) ale, že	+	41,842	15.58
2.	(i)	Řekl (...) však, že	+	50,162	18.68
3.	(ii)	Řekl (...), že ale	-	606	0.23
4.	(ii)	Řekl (...), že však	-	294	0.11
5.	(ii)	Řekl (...), že (...) ale[31]	-	3,594	1.34
6.	(ii)	Řekl (...), že (...) však	-	1,570	0.58
7.	(i)	Řekl ale, že	+	24,477	9.12
8.	(i)	Řekl však, že	+	31,822	11.85
9.	(ii)	Řekl, že ale	-	415	0.15
10.	(ii)	Řekl, že však	-	235	0.10

Table 4.13: Distribution of structures with raised and non-raised contrastive connectives in Czech according to SYN series of the Czech National Corpus

connective to connect two events, states, etc., it seems that **the connective tends to stand as close as possible to the right boundary of the first argument in surface order, that means, in most cases, at the sentence boundary.** So, in this view, it does not "climb" up in the structure, it **moves left** in the linear text order. In sentences with the end-position of the main clause, possibly the attribution clause, the connective would not "climb" that often, since this would mean moving to the right, away from both its arguments. The movement to the right would not support the understanding of the relation; the recipient would have to "wait for the connective".

A certain check of the correctness of our reasoning is that the absolute majority of the query results (or at least those that we could check by hand) indeed concerned verbs of speaking and thinking. There was no constraint in the query on the verb's form or meaning, cf. Footnote 27.

So far, we have documented corpus distributions of the examples (130a) and (130b) from above. Testing our hypothesis further by comparing distributions of the examples with the end-position of the main (attribution) clause, (130c) and (130d), is a much harder task without a very large corpus with syntactic annotation. There might be other possibilities of querying unparsed corpora for testing this tendency in connective behavior, constructing such a query is, however, non-trivial. There is no easy device to reliably set the border of the last clause in a sentence (apart from looking for commas, which is very inaccurate). In the syntactically annotated data of the PDT, we were able to detect only a small number of these connectives, and moreover, the results were mainly such occurrences where the last clause of a sentence is interrupted by an embedded clause and the connective in question appears **before** the interruption. This makes the notion of the last clause of a sentence problematic and shows again the tendency of the connective not to stand too right, too far from its first argument.

As a supplementary, transformation-based test of our claims, we can raise the connective in a rather long sentence with the end-position of the attribution clause. This means, for such a case, moving the connective to the right, away from its first argument. We do not treat the resulting sentences as ungrammatical, but, in our view, such a transformation hinders the comprehensibility of the discourse. Compare (131a) – original sentence and (131b) – sentence with the raised connective *ale*. In (131c) and (131d), the attribution clause is moved to the front, in (131c) the connective is moreover raised (moved to the left); in (131d) it is preserved at its original position within the dependent clause. Given the length of the example, the variants (131b) and (131d) seem quite awkward. In (131b), the contrastive meaning is apparent much earlier than when the connective actually appears. The sentence (131d) is then, at least in Czech, stylistically inapt.[32]

(131a) *Obrovské štěstí je, že je léto a lidé jsou tady ve městě schopni vypěstovat alespoň zeleninu. Hlad zatím nehrozí.* **Zima by ale byla v současných podmínkách katastrofální, protože UNHCR (Vysoký komisariát OSN pro uprchlíky) prostě už nemá peníze na pomoc a možná se stáhne,** říká pracovník humanitární agentury v malém baru v centru Sarajeva.

Enormous luck is the fact that it is summer and people are able to grow at least vegetables here in the city. There is no hunger yet. **But the winter** [lit. The winter but] **would be disastrous in the current conditions because UNHCR (The United Nations Refugee Agency) simply has no money to help and maybe it will withdraw,** says a humanitarian organization employee in a small bar in the center of Sarajevo.

[32] The translations of this example to English are the best approximations to the Czech originals; they can, however, have a different degree of acceptability. We use translations of *ale* both with *but* and *however* to preserve the original word order. The Czech *ale* has more relaxed word order rules than the English *but.*

(131b) *Obrovské štěstí je, že je léto a lidé jsou tady ve městě schopni vypěstovat alespoň zeleninu. Hlad zatím nehrozí.* **Zima by byla v současných podmínkách katastrofální, protože UNHCR (Vysoký komisariát OSN pro uprchlíky) prostě už nemá peníze na pomoc a možná se stáhne**, říká ale pracovník humanitární agentury v malém baru v centru Sarajeva.

Enormous luck is the fact that it is summer and people are able to grow at least vegetables here in the city. There is no hunger yet. **The winter would be disastrous in the current conditions because UNHCR (The United Nations Refugee Agency) simply has no money to help and maybe it will withdraw,** says however a humanitarian organization employee in a small bar in the center of Sarajevo.

(131c) *Obrovské štěstí je, že je léto a lidé jsou tady ve městě schopni vypěstovat alespoň zeleninu. Hlad zatím nehrozí.* Pracovník humanitární agentury v malém baru v centru Sarajeva ale říká, **že zima by byla katastrofální, protože UNHCR (Vysoký komisariát OSN pro uprchlíky) prostě už nemá peníze na pomoc a možná se stáhne.**

Enormous luck is the fact that it is summer and people are able to grow at least vegetables here in the city. There is no hunger yet. A humanitarian organization employee in a small bar in the center of Sarajevo says, however, **that the winter would be disastrous in the current conditions because UNHCR (The United Nations Refugee Agency) simply has no money to help and maybe it will withdraw.**

(131d) ? *Obrovské štěstí je, že je léto a lidé jsou tady ve městě schopni vypěstovat alespoň zeleninu. Hlad zatím nehrozí.* Pracovník humanitární agentury v malém baru v centru Sarajeva říká**, že zima by ale byla katastrofální, protože UNHCR (Vysoký komisariát OSN pro uprchlíky) prostě už nemá peníze na pomoc a možná se stáhne.**

? *Enormous luck is the fact that it is summer and people are able to grow at least vegetables here in the city. There is no hunger yet.* A humanitarian organization employee in a small bar in the center of Sarajevo says **that the winter, however, would be disastrous in the current conditions because UNHCR (The United Nations Refugee Agency) simply has no money to help and maybe it will withdraw.**

The analyses in this subsection have, in our opinion, sufficiently demonstrated motivation for connective raising. This motivation is to achieve coherence, comprehensibility of the discourse and it is so strong, that it has priority over the syntactic rules. For the lack of data in other meaning classes we could only make claims about the contrastives in this respect. We have learned about them that they can be interpreted as "raised" only in such contexts where the contrastive relation is the content of the claim of one source (speaker) only or where a second source, claiming a contrastive statement, is not important for the overall purpose of the given discourse.

4.2.4 Summary

According to our findings, mismatches between syntax and discourse levels of language description are mostly regular and can be systematically described. We have analyzed several most common types of structural mismatches which initially caused inter-annotator agreement decrease. We found out that our annotation scheme for discourse had been originally partly underspecified (in terms of negation treatment etc.), and, that, for a correct semantic interpretation of the discourse – and on its basis a correct data representation for the NLP – an analysis of attribution is urgently needed. This holds especially for phenomena which can be interpreted elsewhere in the text than where they appear in surface form (negation scope and connective "scopes"). As we have verified with the analysis of bigger amount of data in the Czech National Corpus, such mismatches need not always be rare or marginal phenomena, for instance, connective raising proved to be quite a common behavior of connective devices in discourse. Nevertheless, we must admit the existence of a grey area where the linguistic tests accessible to us are not sufficient for a satisfactory interpretation of certain complex discourse phenomena in real data. This is a logical consequence of performing linguistic analyses on such a complex level of language description.[33]

There might be other mismatches on the syntax – discourse interface than those described in this section. We only concentrated here on those that we were able to track down in the available annotated corpus data and that appeared, at least to a certain extent, repeatedly.

[33] For a detailed methodological study on the possibilities and limits of corpus approaches to such complex phenomena like discourse structure or coherence, see Poláková (2014).

4.3 Implicitness in Discourse

The present section has developed in addition to the intended outline, as a consequence of our contemplation on implicit phenomena in discourse. It was motivated by the wish to introduce a unified, well thought-out approach to analyzing and annotating "hidden" signals of discourse coherence for the Prague Treebank. We do not speak here about cognitive processes that enable text understanding in general (more on them in Section 2.4.2 in the theoretical part of this thesis), although they are also "hidden". We speak here, in agreement with the PDTB approach, about discourse relations with no connective device present in the text, cf. Example (132).

(132) *Jane will not come. She is sick.* LP

Describing implicit phenomena in discourse is a problematic task, both theoretically and empirically. Therefore, apart from our own research, we closely observe and analyze experience of the research community with annotating, describing and automatic processing of implicit relations, in the PDTB approach and in other perspectives.

In this section, we first characterize the PDTB approach to the implicits and compare its results to the results in related discourse corpora. Next, we describe our own annotation experiment, a pilot study of implicit relations annotation possibilities, based on the Penn annotation approach and conducted on the Czech corpus data. Last but not least, we offer a survey of the latest efforts by other researchers regarding this topic, pointing out the results of two particularly important studies for our purpose. The first one describes an experiment in making the implicit relations explicit, putting thus the notion of implicit relations into a different perspective (Taboada and Das 2013, Das 2014). The second one aims at an automatic recognition of the implicit relations (Lin et al. 2009). Summing up findings of these annotation projects and experiments, we search for a mutually shared experience that would form a well-founded account of implicit phenomena in discourse.

4.3.1 Implicit Relations in the PDTB and Related Corpora

In the PDTB, the term *implicit relations* represents discourse relations which are not signaled by any discourse connective in the text (on the surface). As already mentioned earlier in Section 2.3.2.2, these relations must be inferred by the recipient from the semantic contents of the text units and from the context. In the PDTB 2.0, the annotation of implicit relations was carried out in the whole corpus (approx. 50,000 sentences). The annotators inserted a connective that best fits to express the inferred discourse relation between all successive pairs of sentences within paragraphs, but also intra-sententially between clauses delimited by semi-colon or colon (Prasad et al. 2008). These inserted expressions are called *implicit connectives*, cf. Example (133) from the PDTB.

(133) *In July, the Environmental Protection Agency imposed a gradual ban on virtually all uses of asbestos.* (implicit = as a result) *By 1997, almost all remaining uses of cancer-causing asbestos will be outlawed.*

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 counted together with the figures for explicits, so the actual statistics on implicit relations remains unpublished.[34] From personal discussions with annotators experienced with this task, it shows that the agreement on the implicits is perceived rather low, far beyond satisfactory. This was also the reason for postponing such a task in the Prague 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.

There were two major questions to answer before initiating such a task: do we really need some (possibly hand-crafted) representation of implicit relations and if yes, how to do it reliably?

The answer to the first question is “yes”: having annotated only the explicit relations, that means those expressed by explicit connectives, has a linguistic value, but for any automated modeling of discourse structure this data is probably too sparse. There may be other phenomena annotated in the data that are relevant for discourse processing, like alternative lexicalizations of the connectives, but the absence of information on implicit relations can be quite a big loss. This is clearly shown at least by the PDTB annotation results: the number of inferred (implicit) relations in the treebank texts is almost equal to the number of the explicit, connective-based (16,224 : 18,459), (Prasad et al. 2008); in Hindi Discourse Relation Bank the ratio is 185 : 189 (Oza et al. 2009); in the Biomedical Discourse Relation Bank, the implicit relations even slightly outnumber the explicit ones 3,001 : 2,636 (Prasad et al. 2011). The LUNA corpus of

[34] To our knowledge, there are only few published inter-annotator agreement figures solely for the implicit relations. In the PDTB, 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. 2008). Agreement on the inserted connectives was 72% (for 5 semantic categories), (Miltsakaki et al. 2004, p. 7).

Italian spoken dialogs, however, shows a different tendency: the ratio of implicit and explicit relations is 487 : 1,052 (Tonelli et al. 2010, p. 2089).[35] This may be caused by a slightly different annotation procedure and by the nature of the spoken dialogical texts themselves.

The second question, namely how to design such a task in order to get reliable results, is a part of a larger discussion in the linguistic community on implicit phenomena in discourse. This discussion grows in the recent years, as the increasing number of publications on this topic shows up. Only the very approach of the Penn group inspired many discourse-oriented researchers to various experiments. The PDTB creators themselves gathered feed-back on their annotations, which they summed up in a recent study by Prasad et al. (2014). Regarding implicit relations, the study points out the consequences of the annotation decisions: Primarily, the consequences of the fact that the category of the implicit relations was established before establishing other categories, which, had they been annotated first, would perhaps reduce the number of the original implicit relations. Originally in the PDTB, where the annotators could not provide an implicit connective between adjacent sentences, they looked for other possible connecting signals. If they felt the connection of the sentences happens through an entity identity (coreference), they inserted the "EntRel" label. If it happens through a larger connecting phrase, they inserted the "AltLex" label (alternative lexicalization of the connective), if they could not provide any of the previously mentioned categories, they inserted a label for no relation "NoRel". Some of these consequences will be addressed also later in this section.

4.3.2 Pilot Annotation Experiment[36]

We, also, have contributed to this feed-back by conducting an experimental annotation of implicit relations on 100 sentences from the PDT. The experiment is shortly described in Poláková et al. (2013), here, we provide a more detailed description. The objective of this pilot annotation was to attempt to achieve a higher inter-annotator agreement by removing factors discovered as repeatedly disturbing. Our experiment was designed following the basic principles of Penn annotation of the implicits, with only slight adjustments:

1. Between every two adjacent sentence units, except for paragraphs boundaries and except for cases where there already was an explicit connective annotated, (that means, in accordance with the PDTB), our annotators put a mark

[35] As far as we know, other projects with annotation of discourse relations do not have completed any annotation of implicit relations so far but there are some in preparation.

[36] We would like to thank Pavlína Jínová for her part in this experiment – for annotation work and for helping with the analysis of the results.

File	No. of explicit r.	No. of implicit r.	Strict IAA (abs.)	Strict IAA (%)	Liberal IAA (abs.)	Liberal IAA (%)
1.	8	15	10	66.7	10	66.7
2.	4	21	9	42.9	12	57.1
3.	2	19	8	42.1	10	52.6
total	**14**	**55**	**27**	**49.1**	**32**	**58.2**

Table 4.14: Strict and liberal IAA on implicit relations in a pilot annotation

for the most appropriate semantic type of discourse relations (not a connective expression, unlike in the PDTB). In addition to the various semantic types of discourse relations, there was also the possibility to mark the connection of the two sentences as being established by other means (the label *other*). These means were then divided further to three sub-labels: coreference-based (the label *other-coref)*, based on associative anaphora (the label *other-bridging*) – these two categories correspond to the Penn label "EntRel", entity based relation – and the *other-NoRel* label for cases, where no relation could be found.[37]

2. According to the PDTB annotators, the Wall Street Journal texts proved to be sometimes difficult to understand by people without an economic background. Therefore, our text types were taken from a domain accessible to the annotators, reviews of cultural events. Also, the chosen texts were short, up to 35 sentences each.

3. The annotation was carried out by two annotators most experienced with the large-scale annotations of explicit connectives.

4. The annotation was carried out on plain texts, not on the syntactic trees, as it is the case in the full-scale annotations of explicit relations in the PDT. The only information available to the annotators was whether there already is an explicit relation annotated in the slots between adjacent sentences or not.

5. The annotation of argument extents of the inserted implicit relations was not conducted in this experiment.

The results are summed up in Tables 4.14 and 4.15. In a strict inter-annotator agreement (IAA) measurement, the annotators agreed in 49.1% on type of the relation, counting the three „other"- subcategories separately (22 + 3 = 25 labels). In a more liberal IAA measurement, where choosing whichever of the *other*-relations was taken as agreement (22 + 1 labels), the agreement was slightly

[37] The „AltLex" category was taken to be a subset of explicit connectives. Their annotation was at the time of the experiment work in progress.

Disagreement pair	Occurrences
conjunction x coreference	7
specification x coreference	3
bridging x coreference	2
NoRel x coreference	2
conjunction x bridging	2

Table 4.15: Most frequent disagreements in implicit relation annotation

higher – 58.2%. The most problematic issue revealed to be distinguishing between the relations from the EXPANSION class (conjunction, specification etc.) on one side and between relations only based on coreference on the other, as demonstrated by Examples (134) and (135).

(134) *Zvláště jímavá je fotografie Burljuka z jeho amerického pobytu, kde stále zdůrazňoval svou futuristickou image.* (implicit: specification) **Jednu tvář má pomalovanou, připomíná tetování, v uchu náušnici, na hlavě cylindr, pruhovanou křiklavou vestu...**

Especially catchy is a photo of Burliuk from his stay in America, where he constantly emphasized his futuristic image. (implicit: specification) **One cheek painted, resembling a tattoo, an earring in one ear, a top hat on his head, a garish striped vest...**

(135) *Někteří si vystoupení dudáckých souborů zaznamenávali na videokamery, případně nahrávali na magnetofony.* (implicit: conjunction/bridging) *Bylo i nebylo co nahrávat.*

Some (spectators) recorded the performances of the bagpipe bands on video cameras or on tape recoders. (implicit: conjunction/bridging) *There was plenty and nothing to record.*

Example (134) was interpreted by both annotators identically as a case of specification, whereas (135) was evaluated as a case of conjunction by one annotator (in the PDTB methodology, the implicit connective *and* would be inserted) and as a case of a bridging anaphora by the other annotator. Table 4.15 shows the distributions of the disagreements that occurred more than once.

The disagreement analysis also indicates that agreement on a semantic type of a discourse relation, if any was detected, was much less problematic (6 out of 28 disagreements in total). Apparently, the annotators were able to agree in most cases after a subsequent joint analysis, as one of the interpretation usually fitted more than the other.

The restriction of the annotation only to slots between adjacent sentences was found useful for a simplification of the annotation. Also, preserving this constraint in further annotation would make the results comparable to those in the PDTB. On the other hand, this adjacency principle did not always match the annotators' intuition where the argument borders should be. This intuition is also supported by the figures for the explicit relations in the PDT 3.0 where 5.6% (or 1,163 relations) have a non-adjacent argument. This was also noticed by the PDTB creators. They have loosened this rule in a later annotation of the English biomedical texts in BioDRB (Prasad et al. 2011) by allowing the annotators to search also for a distantly-placed left argument of an implicit connective. They 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). The Hindi DRB (Kolachina et al. 2012, Oza et al. 2009) and the LUNA corpus of spoken Italian (Tonelli et al. 2010) follow the same strategy in this respect.

In our annotation experiment, we, too, confirmed adequacy of the tendency of new discourse projects to cancel the adjacency criterion for annotating implicits. This makes the task a little more complex but also more linguistically adequate.

On the other hand, we have applied four measures to simplify such a task. The results, however, are still not perceived as an adequate start for a full-extent annotation of Czech texts. The annotation experiment was performed on a small sample of data; any of our findings have therefore only a limited effect. Other types of texts may show other types of difficulties. Further research, both theoretical and empirical, is needed in this field.

4.3.3 Explicitation of the Implicit Relations

A similar method, namely revising the category of implicit relations by relaxing some annotation rules, is used in a study by Taboada and Das (2013) and in the dissertation thesis of Das (2014). Their experiment is particularly worth mentioning here since it redefines the notion of implicit relations and brings a new insight on how coherence means function in general. The authors were interested in the possible ways of signaling a discourse relation, claiming that signaling is very common and that it includes other phenomena than just conjunctions or discourse markers. According to them, there are no (or very few) implicit relations. For testing their hypothesis, Taboada and Das used the tree structure annotations of the RST treebank (385 WSJ articles). This data was originally annotated for discourse (rhetorical) relations only[38] (as a part of the overall discourse structure), that means neither connectives nor other such discourse-connecting devices were annotated. The experiment consisted in looking at all possible kinds of signals that led to identification of a discourse relation, and classifying them. 8 groups of signals were identified (discourse markers, reference, lexical, semantic, morphological, syntactic, graphical, and genre) with 39 types of signals in total. Also, a characteristic of each relation regarding its typical way of signaling was provided. Altogether, 92.74% of relations were signaled either by a discourse marker, by other signals, or by both; the remaining 7.26% is then the actual proportion of unsignaled (implicit?) relations[39]. This is a very small number in comparison to the proportions of implicits in other corpora mentioned earlier in this chapter. The 7.26% of unsignaled relations are further characterized as nonexistent, tenuous or questionable discourse relations (not included in the RST taxonomy), like comment or topic-shift.

The study of Taboada and Das shifts the understanding of implicitness (in the PDTB sense) to a new perspective: the absence of any signal appears to be much rarer in discourse relations. Many of the implicit relations (in the PDTB sense) are thus not implicit at all. A negative definition of implicit relations is offered by saying what they are not. The remaining small proportion of relations without signaling is heterogeneous. The difference between a relation not being signaled (implicit in the new sense) and the non-existence of a relation (NoRel) is not addressed in this study, as the RST analysis presupposes one continuous discourse structure, with all discourse units connected (i.e. NoRel is not possible). For a researcher outside the RST framework, the problem has only shifted to another level: how to distinguish between relations with no signaling and non-existence of the relation itself. So far, we do not have well-founded claims on this issue. The research on NoRels in Czech is a current research topic. In this phase, we can only state that we share the assumption of Taboada that there may be no implicit relations at all[40]. The 7.26% would then correspond to an absence of discourse relations (NoRels: interruptions in the text, topic shifts etc.)

The study is, moreover, especially valuable for its taking into account and describing the different aspects, by which coherence is realized. Similarly as the few highly valued multi-dimensional discourse corpora, the study does so by

[38] Here, discourse relations are equal to rhetorical relations, and can be also referred to as coherence relations.

[39] The numbers reported here differ slightly from those reported in Taboada and Das (2013). We refer here to an updated version of the project presented by M. Taboada at the COST meeting in Louvain-la-Neuve in Jan. 2015. Some more details will follow with the release of the RST Signalling Corpus (Das et al. 2015).

[40] Again, we cite the presentation by M. Taboada in Louvain-la-Neuve, January 2015.

studying authentic texts, by looking which particular language means participate in creating discourse coherence and how. Of course, there can arise the question how to define such groups of signals, and what the discourse-structuring relevance of the individual signals is. The relevance of discourse connectives is high, as the discourse-structuring (or connecting) function is their primary one. With tenses of verbs, for instance, it is much less clear. But even having mapped such signals, whose function might not be in the first place signaling discourse structure, is a huge contribution to the effort of finding out how we really understand.

4.3.4 Human and Automatic Recognition of the Implicit Relations

So far, this section has addressed the issues of human recognition and annotation of the implicit relations. Another direction of research dealing with implicit relations which is worth mentioning here is concerned with automatic recognition of them (Pitler et al. 2009, Zhou et al. 2010, Lin et al. 2009), since it defines and analyzes problematic areas of such a task. The state-of-the-art recognition systems take advantage of various linguistic features and their combination, including contextual modeling of discourse relations, features extracted from constituent parse trees and dependency parse trees, word pair features etc. Lin et al. (2009) improved the majority baseline for classification accuracy by 14.1%, but, as they state, "Although we feel a 14.1 absolute percentage improvement is a solid result, an [overall] accuracy of 40% does not allow downstream NLP applications to trust the output of such a classification system". This shows the difficulty of the task also in automatic processing. More importantly, Lin et al. (2009) further offer an elaborated analysis and classification of difficulties that caused occasional poorer performance of their system. They name four areas as particularly difficult:

1. ambiguity – some relations are very similar to each other in terms of words, syntax and semantics; moreover, some of the connectives can be ambiguous. Also, the PDTB annotators had the choice to assign a double label (e.g. Contrast/ Conjunction) in case of doubt. More context would be needed to disambiguate these relations.

2. inference – In cases like (136), we need to decide whether the semantics of the first argument infers that of the second or the other way round. A formal semantic representation of each relation, a knowledge base, would be needed in order to understand these cases.

(136) *I had calls all night long. I was woken up every hour.*

3. contextual modeling – sometimes, the two discourse arguments are not enough to understand the meaning of the relation, a wider context (possibly also of the whole text) is needed.

4. world knowledge – the authors point out the importance of further world knowledge, not only the knowledge of the context. They show it on an example, where understanding of the metaphor of “Trojan Horse“ as a gift with harmful intent is essential for understanding the correct meaning of the relation.

The study is summed up with the ascertainment that “implicit discourse relation classification needs deeper semantic representations, more robust system design, and access to more external knowledge” (Lin et al. 2009, p. 350). It is also pointed out, that these findings may not be relevant to only recognizing implicit relations, but also apply to other discourse-related tasks.

Comparing the results of our annotation experiment with the analysis of Lin et al. (2009) brings us to the observation that it is exactly the same areas of difficulties that cause human annotators to hesitate when annotating implicit relations, or even disagree. Understanding the context is crucial. But even a skilled annotator can have difficulties when it comes to a semantic interpretation of very complex texts, or texts from domains that the annotator is not familiar with – where he or she, just like the recognition system, lacks domain knowledge.

4.3.5 Summary

This section aims at getting a broader insight into the issues of implicit discourse relations in order to prepare such an annotation task for Czech language data. Based on our own annotation experiment and on recent findings to this topic in international discourse-oriented research, we have arrived at the observations listed here and further discussed below:

1. There are more notions referred by the term “implicit” relations in discourse. The largely followed PDTB approach uses the term for absence of discourse connectives. The introduced studies of Taboada and Das (2013) and Das (2014) consider signaled and non-signaled relations. Here, not only the absence of an explicit discourse connective but the absence of any signal of coherence at all is referred to as an “implicit” relation. In both cases, the definition of a “connective” and that of a “discourse signal”, respectively, plays a role for determining what an implicit relation is.

2. We share the opinion of Taboada and Das (2013) that discourse relations in fact in the vast majority have some sort if signaling, even though is it not explicit connectives or their alternative lexicalizations.

3. The reason why annotating implicit relations (and, as a second step, their automatic recognition) is difficult, is that it requires a fair portion of contextual

information (the co-text, situational context, domain/world knowledge). In authentic texts, moreover, one encounters vague continuation semantics, weak or ambiguous relations and domain-specific interpretation problems.
Consequences:
There should be a careful distinction between implicit relations and implicit connectives. Whereas the existence of implicit relations in the PDTB sense (not signaled by a discourse connective) is undeniable, the insertion of connectives between arguments of such implicit relations is an annotation-based decision, with the connective being a substitute for the semantic category of the relation. The PDTB annotation procedure, which allows to mark Altlex, EntRel or NoRel (as explained above) in case where neither explicit nor implicit connective fits in, supports the Taboada and Das (2013) claim that there can be other signals of coherence. In fact, 37% of discourse relations in the RST annotation project have multiple signaling. If we want to know the semantic type of the relation, annotating the semantic category is enough (even though it may be harder for the annotators). If we want to know how the relation is signaled on the surface, we can look for these less apparent signals and annotate them as markers of “non-connective” relations.

The annotation decision to insert an implicit connective has also other aspects: Our annotation experience is that sometimes the insertion of a connective seems clear (supposedly when contrasting two facts or implying one from the other) as shown on Example (132) above, the perceived semantic relation is strong. But sometimes the insertion of a connective seems forced, mostly where there is a weak or unclear semantics of addition, continuation or restatement (specification, equivalence etc.), as demonstrated by Examples (135) and (136). Also, from the point of view of the recipient, inserting of a constructed connective might get confusing if we started thinking that a connective is necessary for understanding the relation. It is quite the opposite: if there is no explicit connective, it does not have to be there – we can understand nevertheless.

For a corpus like the Prague Dependency Treebank, in which an extended annotation of textual coreference is finished (Nedoluzhko 2011) and annotation of Altlexes (or secondary connectives) is on the way (Rysová 2012a, 2012b), a different course of annotation seems more convenient – one that would profit from the existing annotations. Also, we suggest taking into consideration the recent view of Prasad et al. (2014) that implicit relations (in the PDTB sense) can be found between other texts spans than only between adjacent units. Another suggestion for further advancement would be not to attempt a full coverage (between every two sentences) for implicit semantic relations in manual annotation. It is necessary to accept the fact that in complicated texts discourse relations are often vague, ambiguous or weak and the analysts tend to disagree in their interpretations. Still, it would be worth to have covered places which can be interpreted uniformly and reliably.

Moreover, an interesting point here would be to examine whether the possibility to insert an implicit connective between certain discourse units relates in some way to the presence or absence of other signals of coherence, to their types and strength as connecting devices. Such a comparative study even suggests itself for English, as the texts of the RST Signalling Corpus (Das et al. 2015) and the PDTB are identical.

Computationally, such an annotation could offer (i) a higher reliability in gold standard data (ii) a linguistically more adequate information about the true signals of coherence for relations with no explicit connectives/altlex, and about the location of the arguments.

5

Discussion and Conclusions

5.1 Discussion

In this chapter, we summarize the most important issues that have arisen during our research. In the theoretical part of the thesis, we focused mostly on the properties assigned to discourse structure and discourse relations in various concepts of discourse analysis and we discussed the possibilities of an adequate representation of these properties in a corpus annotation. The approaches on the international scene mostly agree on the following assumptions about discourse: (i) a coherent discourse is interconnected: it has some type of coherence among all of its segments; (ii) discourse relations, i.e. semantic relations between propositions, relate two such segments and, typically, there is some kind of signal of this relation (not necessarily a discourse connective); (iii) discourse relations are in some way closely connected to syntactic structure. Even the models with the "global" modeling of discourse make use of syntactic features. Moreover (iv), at least some semantic categories that apply in syntax and in discourse are identical. Also, many researchers addressed the issue of where the source of coherence comes from: (v) discourse relations hold not only between the very contents of the discourse segments, they can also relate unexpressed contents (illocutions, inferences etc.).

All these assumptions were taken into consideration (whereas others were not – like the representation of a whole discourse by a tree graph) and are reflected in the proposal of our annotation scheme for discourse relations in Czech. Although our scheme for analysis of discourse relations is a first such attempt of its kind for Czech, we can claim that our model is in general features stable and functioning – we can document this by the satisfactory level of consistency in all aspects of the annotations.

Furthermore, one of the aims of the present study was to reflect on the pros and cons of the proposed way of analysis. The real touch with the Czech data indicated some weak points, the most important of which were addressed in the third part of the thesis. The difficulties during the annotation process and also the evaluation results lead us to the following convictions:

(i) An analysis of attribution, i.e. the distinction of clauses of reporting and the reported content, is a necessary annotation component. Also, one has to remember that certain phenomena (negation, contrastive connectives) in connection with attribution can be interpreted somewhere else in the structure than where they appear on the surface. Not having included this knowledge in the annotation caused confusion regarding the delimitation of a discourse argument, or more precisely, the extent of the argument.

(ii) A semantic categorization of discourse relations must not ignore the "pragmatic" factors. Major disagreements in assigning semantic types to the relations were between types where a pragmatic (subjective) component plays a substantial role. The highest disagreement in the PDiT 1.0 was with the distinctions *reason – result* vs. *explication* and *opposition* vs. *concession.* This implies that either these categories were defined too vaguely (the former case, in our view) or that there is general disagreement in what kind of inferences are there behind these meanings (the latter case). The concessive meaning is established by a denial of an expectation – the problem in the annotations was caused most likely by the different views on whether there is any expectation (shared by the author and the recipient) at all.

Also, there are very few pragmatic types of relations annotated. This can be due to the nature of the data domain (informative journalistic texts), but also due to the lack of attention for the subjective – objective ("semantic" – "pragmatic") distinction. For any future extension of the semantic classification, apart from other small changes, we propose a three-way division of the relations concerning their source of coherence: semantic (content), epistemic and speech act.

(iii) The large-scale analysis of authentic texts turned our attention not only to more subtle methodological questions concerning an adequate and reliable representation of discourse relations but also to the limits of such a representation. In our experience, the borders of what can be reliably described and agreed on in authentic texts when analyzing language phenomena with such a high degree of complexity are dependent on the explicitness of the language means used. In our view, there is little agreement in labeling discourse semantics without labeling the surface forms of its expression. We have documented this on the example of implicit discourse relations (and connectives) which show low agreement values when annotated by humans and low accuracy when modeled by computers. To capture discourse relations with no explicit discourse connectives, we propose to concentrate on other surface signals anchoring the relations like referential expressions, lexical and morphological signals, specific principles in syntactic structure etc.

(iv) The PDiT 1.0 annotation scheme is probably the only discourse-oriented annotation project making direct use of the existing syntactic annotation. The tectogrammatical trees with resolved syntactic structure of a sentence turned out to be of great advantage for the discourse analysis (verb ellipsis restoration, easy extraction of intra-sentential relations, coreference annotation accessible). While there already is an extensive research on the commonalities of syntactic structure of a sentence and the structure of a discourse, our method enabled us to focus also on their discrepancies. For instance, we were able to easily detect and analyze cases where discourse arguments did not correspond to clauses or sentences and describe why this happens. In addition, the principle of interlinking the analyses "within a sentence" (morphology, surface syntax, underlying syntax with information structure, multiword entities etc.) and "beyond the sentence boundary" (discourse relations, extended coreference annotation, bridging anaphora, genre distinction etc.) in the PDT 3.0 offers its users a unique resource of multifarious linguistic information accessible and searchable within a single representation.

5.2 Conclusions

In the present thesis, we have proposed a complex linguistic scheme for analysis of discourse relations in Czech, and, for the first time in Czech linguistics, we have attempted their formalized representation in a large-scale language corpus (approx. 50 thousand sentences). We have described all phases of the annotation process, evaluated the annotation accuracy, provided a quantitative overview of the annotation results and offered a linguistic analysis of some of the problematic issues in the annotation.

The resulting language resource, the Prague Discourse Treebank 1.0, was released in 2012 and its updated and extended version, the Prague Dependency Treebank 3.0, appeared one year later. Both corpora, along with the viewing and searching interface, are freely available in the LINDAT-CLARIN repository.

The main aim of this thesis was to contribute to the general knowledge about discourse coherence by investigation of discourse relations with methods of corpus linguistics on Czech language material. We believe that our study fulfils this aim, in particular by contributing to our understanding of discourse structure and coherence in the following three aspects:

First, the systematic description of discourse relations embodied in a searchable corpus builds a solid base for any linguistic research in the area of discourse analysis.

Second, it offers gold standard data for computational experiments and applications ranging from connective disambiguation, across automatic text analysis and summarization, to the solution of issues of coherence in machine translation.

Last but not least, our study demonstrates a transfer of a methodological concept, or more precisely, it documents the application of a framework originally developed for English (the PDTB approach to discourse relations) on a typologically different language. In this way, we offer feedback to the authors of the original concept and indirectly contribute to the topic of the role of language universality/specificity in linguistic methodology.

Bibliography

Adamec, P. (1995). Konektivní částice a jiné textově propojovací výrazy v současné češtině. In *Přednášky z 37. a 38. běhu LŠSS*, pages 59–64, Praha. Univerzita Karlova.

Afantenos, S., Asher, N., Benamara, F., Bras, M., Fabre, C., Ho-Dac, L.-M., Le Draoulec, A., Muller, P., Péry-Woodley, M.-P., and Prévot, L. (2012). An empirical resource for discovering cognitive principles of discourse organisation: the ANNODIS corpus. In *Proceedings of LREC 2012*, pages 2727–2734, Istanbul, Turkey.

Al-Saif, A. and Markert, K. (2010). The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic. In *Proceedings of LREC 2010*, pages 2046–2053, Valletta, Malta.

Asher, N. (1993). *Reference to abstract objects in discourse.* Kluwer, Norwell, MA.

Asher, N. and Lascarides, A. (2003). *Logics of conversation.* Cambridge University Press.

Bedřichová, Z. (2008). Částice implikující presupozici v češtině. Master's thesis, Faculty of Philosophy and Arts, Charles University in Prague, Czech Republic.

Bejček, E., Hajičová, E., Hajič, J., Jínová, P., Kettnerová, V., Kolářová, V., Mikulová, M., Mírovský, J., Nedoluzhko, A., Panevová, J., Poláková, L., Ševčíková, M., Štěpánek, J., and Zikánová, Š. (2013). *Prague Dependency Treebank 3.0.* Univerzita Karlova v Praze, MFF, ÚFAL, Prague, Czech Republic.

Bejček, E., Panevová, J., Popelka, J., Smejkalová, L., Straňák, P., Ševčíková, M., Štěpánek, J., Toman, J., Žabokrtský, Z., and Hajič, J. (2011). *Prague Dependency Treebank 2.5.* Univerzita Karlova v Praze, MFF, ÚFAL, Prague, Czech Republic.

Carlson, L. and Marcu, D. (2001). Discourse tagging reference manual. Technical report.

Carlson, L., Okurowski, M. E., Marcu, D., et al. (2002). *RST Discourse Treebank.* Linguistic Data Consortium, University of Pennsylvania, Philadelphia.

Daneš, F. (1968). Typy tematických posloupností v textu. *Slovo a slovesnost*, 29(2):125–141.

Daneš, F. (1985). *Věta a text.* Academia, Praha.

Daneš, F., Grepl, M., and Hlavsa, M. (1987). *Mluvnice češtiny III. Skladba.* Academia, Praha.

Danlos, L., Antolinos-Basso, D., Braud, C., and Roze, C. (2012). Vers le FDTB: French Discourse Tree Bank. In *TALN 2012: 19ème conférence sur le Traitement Automatique des Langues Naturelles*, pages 471–478.

Das, D. (2014). *Signalling of Coherence Relations in Discourse.* PhD thesis, Simon Fraser University, Burnaby, Canada.

Das, D., Taboada, M., and McFetridge, P. (2015). *RST Signalling Corpus.* Linguistic Data Consortium, University of Pennsylvania, Philadelphia.

de Beaugrande, R.-A. and Dressler, W. (1981). *Introduction to text linguistics.* Longman, London.

Dinesh, N., Lee, A., Miltsakaki, E., Prasad, R., Joshi, A., and Webber, B. (2005). Attribution and the (non-) alignment of syntactic and discourse arguments of connectives. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 29–36. Association for Computational Linguistics.

Dressler, W. (1972). *Einführung in die Textlinguistik.* Niemeyer, Tübingen.

Fairclough, N. (1989). *Language and Power.* Longman, London.

Fillmore, C. J. (1963). The position of embedding transformations in a grammar. *Word – Journal of the International Linguistic Association*, 19(2):208–231.

Firbas, J. (1974). Some aspects of the Czechoslovak approach to problems of functional sentence perspective. *Papers on functional sentence perspective*, 1:11–37.

Gastel, A., Schulze, S., Versley, Y., and Hinrichs, E. (2011). Annotation of explicit and implicit discourse relations in the TüBa-D/Z treebank. *Jahrestagung der Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL 2011).*

Grepl, M. and Karlík, P. (1986). *Skladba češtiny.* Státní pedagogické nakladatelství, Praha.

Grosz, B. J. and Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Grosz, B. J., Weinstein, S., and Joshi, A. K. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Hajičová, E. (1993). *Issues of sentence structure and discourse patterns.* Charles University, Prague.

Halliday, M. A. K. and Hasan, R. (1976). *Cohesion in English.* Longman, London.

Hausenblas, K. (1964). On the characterization and classification of discourses. *Travaux linguistiques de Prague*, 1:67–84.

Hausenblas, K. (1971). *Výstavba jazykových projevů a styl.* Universita Karlova, Praha.

Hnátková, M., Křen, M., Procházka, P., and Skoumalová, H. (2014). The SYN-series corpora of written Czech. In *Proceedings of LREC 2014*, pages 160–164, Reykjavik, Iceland.

Hobbs, J. R. (1979). Coherence and coreference. *Cognitive science*, 3(1):67–90.

Hoffmannová, J. (1983). *Sémantické a pragmatické aspekty koherence textu.* Ústav pro jazyk český ČSAV, Praha.

Hoffmannová, J. (1984). Typen der Konnektoren und deren Anteil an der Organisierung des Textes. *Text and the Pragmatic Aspects of Language*, 1:23–39.

Hoffmannová, J. (1986). Temporální sémantika a text. *Teoretické otázky jazykovědy. Linguistica XVI.*, pages 160–191.

Hoffmannová, J. (1993). Koherence, koheze, konexe...? *Slovo a slovesnost*, 54(1):58–64.

Hovy, E. H. (1990). Parsimonious and profligate approaches to the question of discourse structure relations. In *Proceedings of the 5th international workshop on Natural Language Generation*, pages 128–136.

Hrbáček, J. (1994). *Nárys textové syntaxe spisovné češtiny.* Trizonia, Praha.

Hume, D. (1748). *Philosophical essays concerning human understanding.* A. Millar, London.

Iruskieta, M., Aranzabe, M. J., de Ilarraza, A. D., Gonzalez, I., Lersundi, M., and de la Calle, O. L. (2013). The RST Basque TreeBank: an online search interface to check rhetorical relations. In *4th Workshop "RST and Discourse Studies"*, pages 21–23, Fortaleza, Brasil.

Jakobson, R. (1971). Shifters, verbal categories, and the Russian verb. In *Word and Language*, pages 130–147. Mouton, Paris.

Jínová, P. (2011). Vybrané problematické aspekty konektivních prostředků v rámci anotace mezivýpovědních významových vztahů v PDT. *Bohemica Olomucensia*, 2:138–147.

Jínová, P., Mírovský, J., and Poláková, L. (2012). Semi-Automatic Annotation of Intra-Sentential Discourse Relations in PDT. In *Proceedings of COLING 2012*, pages 43–58, Mumbai, India.

Jínová, P., Poláková, L., and Mírovský, J. (2013). Subordinators with Elaborative Meanings in Czech and English. In *Proceedings of the Second International Conference on Dependency Linguistics (DEPLING 2013)*, pages 128–136, Praha, Czech Republic.

Jínová, P., Poláková, L., and Mírovský, J. (2014). *Sentence Structure and Discourse Structure (Possible parallels)*, volume 215 of *Linguistics Today*, pages 53–74. John Benjamins Publishing Company, Amsterdam, The Netherlands.

Karlík, P., Nekula, M., Pleskalová, J., Dočekal, M., Grepl, M., Hladká, Z., Jelínek, M., Křístek, M., Osolsobě, K., Rusínová, Z., Šlosar, D., and Vykypělová, T. (2002). *Encyklopedický slovník češtiny.* Nakladatelství Lidové noviny, Praha.

Kehler, A. (2002). *Coherence, Reference, and the Theory of Grammar.* CSLI Publications, Stanford.

Kim, S.-M. and Hovy, E. (2006). Identifying and analyzing judgment opinions. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 200–207.

Kolachina, S., Prasad, R., Sharma, D. M., and Joshi, A. K. (2012). Evaluation of Discourse Relation Annotation in the Hindi Discourse Relation Bank. In *Proceedings of LREC 2012*, pages 823–828, Istanbul, Turkey.

Kolářová, I. (1998). Významy a funkce slova tedy (teda) v souvislých projevech. *Naše řeč*, 81(2–3):118–123.

Kolářová, I. (2002). Takže jako navazovací výraz v souvětích i souvislých textech a jeho synonymie s výrazy jinými. *Naše řeč*, 85(2):90–97.

Kořenský, J. (1992). *Komunikace a čeština.* H & H, Jinočany.

Lasnik, H. (1976). *Analyses of Negation in English.* Indiana University Linguistics Club, Bloomington.

Lee, A., Prasad, R., Joshi, A., Dinesh, N., and Webber, B. (2006). Complexity of dependencies in discourse: Are dependencies in discourse more complex than in syntax? In *Proceedings of the 5th International Workshop on Treebanks and Linguistic Theories*, pages 79–90, Prague, Czech Republic.

Lin, Z., Kan, M.-Y., and Ng, H. T. (2009). Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 343–351.

Mann, W. C. and Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8:243–281.

Martin, J. R. (1992). *English text: System and structure*. John Benjamins Publishing, Amsterdam, The Netherlands.

Mathesius, V. (1939). O tak zvaném aktuálním členění větném. *Slovo a slovesnost*, 5(4):171–174.

Mathesius, V. (1947). Jazykozpytné poznámky k řečnické výstavbě souvislého výkladu. *Čeština a obecný jazykozpyt*, 1:380–414.

Matthiessen, C. and Thompson, S. A. (1988). The structure of discourse and subordination. *Clause combining in grammar and discourse*, 18:275–329.

Meyer, T. and Poláková, L. (2013). Machine translation with many manually labeled discourse connectives. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Workshop on Discourse in Machine Translation*, pages 43–50, Sofia, Bulgaria.

Mikulová, M., Bejček, E., Mírovský, J., Nedoluzhko, A., Panevová, J., Poláková, L., Straňák, P., Ševčíková, M., and Žabokrtský, Z. (2013). From PDT 2.0 to PDT 3.0 (Modifications and Complements). Technical Report ÚFAL TR-2013-54, ÚFAL MFF UK.

Mikulová, M., Bémová, A., Hajič, J., Hajičová, E., Havelka, J., Kolářová, V., Lopatková, M., Pajas, P., Panevová, J., Razímová, M., Sgall, P., Štěpánek, J., Urešová, Z., Veselá, K., Žabokrtský, Z., and Kučová, L. (2005). Anotace na tektogramatické rovině Pražského závislostního korpusu. Anotátorská příručka. Technical Report TR-2005-28.

Mikulová, M. et al. (2006). Annotation on the tectogrammatical layer in the Prague Dependency Treebank. The Annotation Guidelines. Technical report, UFAL MFF, Prague. Available at: http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/html/index.html.

Miltsakaki, E., Prasad, R., Joshi, A., and Webber, B. (2004). Annotating discourse connectives and their arguments. *Proceedings of the HLT/NAACL Workshop on Frontiers in Corpus Annotation*, pages 9–16.

Miltsakaki, E., Robaldo, L., Lee, A., and Joshi, A. (2008). Sense annotation in the Penn Discourse Treebank. In *Computational Linguistics and Intelligent Text Processing*, pages 275–286. Springer.

Mírovský, J., Jínová, P., and Poláková, L. (2012). Does Tectogrammatics Help the Annotation of Discourse? In *Proceedings of COLING 2012*, pages 853–862, Mumbai, India.

Mírovský, J., Mladová, L., and Žabokrtský, Z. (2010). Annotation Tool for Discourse in PDT. In *Proceedings of COLING 2010*, pages 9–12, Beijing, China.

Mistrík, J. (1975). Skladba textu. *Slovenský jazyk a literatúra v škole*, 22:209–214.

Mladová, L. (2008a). Od hloubkové struktury věty k diskurzním vztahům (diskurzní vztahy v češtině a jejich zachycení v anotovaném korpusu). Master's thesis, Faculty of Philosophy and Arts, Charles University in Prague, Czech Republic.

Mladová, L. (2008b). K problematice vztahu rematizátorů a textových konektorů. *Čeština doma a ve světě*, 16(3-4):126–133.

Mladová, L., Zikánová, Š., Bedřichová, Z., and Hajičová, E. (2009). Towards a discourse corpus of Czech. In *Proceedings of the fifth Corpus Linguistics Conference*, pages 1–8, Liverpool, UK.

Nedoluzhko, A. (2011). *Rozšířená textová koreference a asociační anafora (Koncepce anotace českých dat v Pražském závislostním korpusu).* Studies in Computational and Theoretical Linguistics. ÚFAL, Prague, Czech Republic.

Oza, U., Prasad, R., Kolachina, S., Sharma, D. M., and Joshi, A. (2009). The Hindi Discourse Relation Bank. In *Proceedings of the third Linguistic Annotation Workshop*, pages 158–161.

Pajas, P. and Štěpánek, J. (2008). Recent Advances in a Feature-Rich Framework for Treebank Annotation. In *Proceedings of COLING 2008*, pages 673–680, Manchester, UK.

Panevová, J. (1980). *Formy a funkce ve stavbě české věty.* Academia, Praha.

Pešek, O. (2011). *Argumentativní konektory v současné francouzštině a češtině. Systémové srovnání a analýza okurenční respondence.* Acta Philologica Universitatis Bohemiae Meridionalis, České Budějovice.

Pitler, E., Louis, A., and Nenkova, A. (2009). Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2*, pages 683–691.

Poláková, L. (2014). K možnostem korpusového zpracování nadvětných jevů. *Naše řeč*, 97(4–5):241–258.

Poláková, L., Jínová, P., and Mírovský, J. (2012a). Interplay of coreference and discourse relations: Discourse connectives with a referential component. In *Proceedings of LREC 2012)*, pages 146–153, Istanbul, Turkey.

Poláková, L., Jínová, P., and Mírovský, J. (2014). Genres in the Prague Discourse Treebank. In *Proceedings of LREC 2014*, pages 1320–1326, Reykjavik, Iceland.

Poláková, L., Jínová, P., Zikánová, Š., Bedřichová, Z., Mírovský, J., Rysová, M., Zdeňková, J., Pavlíková, V., and Hajičová, E. (2012b). Manual for Annotation of Discourse Relations in Prague Dependency Treebank. Technical Report 47, Institute of Formal and Applied Linguistics, Charles University in Prague, Prague, Czech Republic.

Poláková, L., Jínová, P., Zikánová, Š., Hajičová, E., Mírovský, J., Nedoluzhko, A., Rysová, M., Pavlíková, V., Zdeňková, J., Pergler, J., and Ocelák, R. (2012c). *Prague Discourse Treebank 1.0.* Univerzita Karlova v Praze, MFF, ÚFAL, Prague, Czech Republic.

Poláková, L., Mírovský, J., Nedoluzhko, A., Jínová, P., Zikánová, Š., and Hajičová, E. (2013). Introducing the Prague Discourse Treebank 1.0. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pages 91–99, Nagoya, Japan.

Polanyi, L. (1988). A formal model of the structure of discourse. *Journal of pragmatics*, 12:601–638.

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 LREC 2008*, pages 2961–2968, Marrakech, Morocco.

Prasad, R., McRoy, S., Frid, N., Joshi, A., and Yu, H. (2011). The Biomedical Discourse Relation Bank. *BMC bioinformatics*, 12:188–205.

Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A., Robaldo, L., and Webber, B. L. (2007). The Penn Discourse Treebank 2.0 Annotation Manual. Technical report, University of Pennsylvania, Philadelphia.

Prasad, R., Webber, B., and Joshi, A. (2014). Reflections on the Penn Discourse Treebank, comparable corpora, and complementary annotation. *Computational Linguistics*, 40:921–950.

Quirk, R., Crystal, D., and Education, P. (2004). *A comprehensive grammar of the English language.* Longman, London.

Robaldo, L., Miltsakaki, E., and Bianchini, A. (2010). Corpus-based Semantics of Concession: Where do Expectations Come from? In *Proceedings of LREC 2010*, pages 3593–3600, Valletta, Malta.

Rysová, M. (2012a). Alternative Lexicalizations of Discourse Connectives in Czech. In *Proceedings of LREC 2012*, pages 2800–2807, Istanbul, Turkey.

Rysová, M. (2012b). Alternativní vyjádření konektorů v češtině. Master's thesis, Faculty of Philosophy and Arts, Charles University in Prague, Czech Republic.

Rysová, M. (2013). K explikativním vztahům v češtině. In *Grenzüberschreitungen – Polnische, tschechische und deutsche Sprache, Literatur und Kultur. Beiträge zur VIII. Internationalen Westslawistischen interFaces-Konferenz in Leipzig*, pages 331–342, Hildesheim, Germany.

Rysová, M. (2014). Verbs of Saying with a Textual Connecting Function in the Prague Discourse Treebank. In *Proceedings of LREC 2014*, pages 930–935, Reykjavik, Iceland.

Rysová, M. and Rysová, K. (2014). The centre and periphery of discourse connectives. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*, pages 452–459, Bangkok, Thailand.

Šaldová, P. (2002). Cohesive devices in newswriting. *Prague Studies in English*, 23:187–200.

Samet, J. and Schank, R. (1984). Coherence and connectivity. *Linguistics and philosophy*, 7:57–82.

Sanders, T. (1997). Semantic and pragmatic sources of coherence: On the categorization of coherence relations in context. *Discourse processes*, 24(1):119–147.

Sanders, T., Spooren, W., and Noordman, L. (1992). Toward a taxonomy of coherence relations. *Discourse processes*, 15:1–35.

Schiffrin, D. (1994). *Approaches to discourse.* Blackwell Publishing, Cambridge.

Sgall, P. (1967). *Generativní popis jazyka a česká deklinace.* Academia, Praha.

Sgall, P., Hajičová, E., and Benešová, E. (1973). *Topic, focus and generative semantics.* Scriptor Verlag.

Sgall, P., Hajičová, E., and Panevová, J. (1986). *The meaning of the sentence in its semantic and pragmatic aspects.* Academia, Prague.

Sgall, P., Nebeský, L., Goralčíková, A., and Hajičová, E. (1969). *A Functional Approach to Syntax in Generative Description of Language.* American Elsevier Pub. Co., New York.

Stede, M. (2004). The Potsdam Commentary Corpus. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, pages 96–102.

Stede, M. (2008). Disambiguating rhetorical structure. *Research on Language and Computation*, 6(3-4):311–332.

Stede, M. and Neumann, A. (2014). Potsdam Commentary Corpus 2.0: Annotation for Discourse Research. In *Proceedings of LREC 2014*, pages 925–929, Reykjavik, Iceland.

Štepánek, J. and Pajas, P. (2010). Querying diverse treebanks in a uniform way. In *Proceedings of LREC 2010*, pages 1828–1835, Valletta, Malta.

Svoboda, K. F. (1956). O souřadných souvětích vysvětlovacích a důsledkových. *Naše řeč*, 39(1–2):1–18.

Sweetser, E. (1991). *From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure*. Cambridge University Press, UK.

Taboada, M. and Das, D. (2013). Annotation upon annotation: Adding signalling information to a corpus of discourse relations. *Dialogue and Discourse*, 4:249–281.

Taboada, M. and Mann, W. C. (2006a). Applications of Rhetorical Structure Theory. *Discourse studies*, 8:567–588.

Taboada, M. and Mann, W. C. (2006b). Rhetorical Structure Theory: Looking back and moving ahead. *Discourse studies*, 8:423–459.

Tárnyiková, J. (2002). *From Text to Texture*. Univerzita Palackého, Olomouc.

Tonelli, S., Riccardi, G., Prasad, R., and Joshi, A. K. (2010). Annotation of Discourse Relations for Conversational Spoken Dialogs. In *Proceedings of LREC 2010*, pages 2084–2090, Valletta, Malta.

Tovena, L. M. (2001). Neg-raising: negation as failure. *Perspectives on negation and polarity items*, 40:331–356.

Versley, Y. and Gastel, A. (2013). Linguistic Tests for Discourse Relations in the Tüba-D/Z Corpus of Written German. *Dialogue and Discourse*, 2:142–173.

Webber, B., Knott, A., and Joshi, A. (1999). Multiple Discourse Connectives in a Lexicalized Grammar for Discourse. In *Proceedings of the Third International Workshop on Computational Semantics*, Tilburg, The Netherlands.

Webber, B., Stone, M., Joshi, A., and Knott, A. (2003). Anaphora and discourse structure. *Computational Linguistics*, 29(4):545–587.

Wolf, F. and Gibson, E. (2005). Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2):249–287.

Wolf, F., Gibson, E., Fischer, A., and Knight, M. (2005). *Discourse Graphbank*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia.

Zeyrek, D., Demirşahin, I., Sevdik-Çalli, A., Balaban, H. Ö., Yalçinkaya, İ., and Turan, Ü. D. (2010). The annotation scheme of the Turkish Discourse Bank and an evaluation of inconsistent annotations. In *Proceedings of the fourth Linguistic Annotation Workshop*, pages 282–289.

Zhou, Y. and Xue, N. (2012). PDTB-style discourse annotation of Chinese text. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 69–77.

Zhou, Z.-M., Xu, Y., Niu, Z.-Y., Lan, M., Su, J., and Tan, C. L. (2010). Predicting discourse connectives for implicit discourse relation recognition. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1507–1514.

Zikánová, Š., Mladová, L., Mírovský, J., and Jínová, P. (2010). Typical cases of annotators' disagreement in discourse annotations in Prague Dependency Treebank. In *Proceedings of LREC 2010*, pages 2002–2006, Valletta, Malta.

Zikánová, Š., Poláková, L., Jínová, P., Nedoluzhko, A., Rysová, M., Mírovský, J., and Hajičová, E. (2015, in press). Zachycení výstavby textu v Pražském závislostním korpusu. *Slovo a slovesnost.*

List of Figures

List of Tables

List of Abbreviations and Acronyms

Altlex – Alternative lexicalization of a connective (in the PDTB scheme)
AO – abstract object
Arg1 – Argument 1 of a discourse relation
Arg2 – Argument 2 of a discourse relation
CDA – critical discourse analysis
CNC – Czech National Corpus (Český národní korpus)
CR – connective raising
DC – discourse connective
EntRel – entity based relation (in the PDTB annotation scheme)
FGD – Functional generative description
HDRB – Hindi Discourse Relation Bank
IAA – inter-annotator agreement
NLP – natural language processing
NoRel – no relation (in the PDTB annotation scheme)
PDiT – Prague Discourse Treebank
PDT – Prague Dependency Treebank
PDTB – Penn Discourse Treebank
PML – Prague Markup Language
PML-TQ – Prague Markup Language Tree Query
PoS – part-of-speech
(S)DRT – Segmented Discourse Representation Theory
SYN – a corpus series of synchronous Czech within the Czech National Corpus
RST – Rhetorical Structure Theory
TR – tectogrammatical representation
TrEd – Tree Editor
WSJ – Wall Street Journal

Appendices

Appendix 1: Semantic classification of discourse relations in the PDiT 1.0 and PDT 3.0 with examples. Arg1 is the text span highlighted in *italics*, Arg2 is highlighted in **bold**. The connective is underlined.

Relation type	Czech example (PDiT 1.0)	English translation
TEMPORAL RELATIONS		
Asynchrony (Precedence – Succession)	**Veškerý vliv nynějšího předsedy ČSSD vyšel v tu chvíli naprázdno** a posléze *zklamal i jeho pokus výsledky pražského sjezdu anulovat.*	**All the influence of the current ČSSD chairman proved fruitless at that moment** and later *his attempt to invalidate the results of the Prague congress failed, too.*
	Štaidl s pomocí detektivní agentury vypátral zmizelou zpěvačku teprve po dvou týdnech. Předtím **mu anonym telefonicky sdělil, že byla unesena.**	*With a help of a detective agency, Štaidl only tracked down the disappeared singer after two weeks.* Before that, **an anonymous call informed him that she had been kidnapped.**
Synchrony (Simultaneity)	*Město postihla krize a nezaměstnanost.* Zároveň **začala nová éra svobodných celních zón.**	*The city was affected by crisis and unemployment.* At the same time, **a new era of free customs zones started.**
	P. Dvorský zahájí program áriemi od B. Smetany a A. Dvořáka. *K této literatuře se hlásím jako k vlastní, řekl Dvorský.* **Zároveň připomněl, že v Čechách se mu vždy dostávalo velké pozornosti.**	P. Dvorský will start the program with arias by B. Smetana and A. Dvořák. *I accept this literature as my own, Dvorský said.* **He** also [lit. *at the same time*] **noted that, in Bohemia, he had always received a great attention.**
CONTINGENCY RELATIONS		
Reason – Result	**Pivo, o jehož názvu by se mělo rozhodnout v průběhu tohoto týdne, je podle jeho slov vhodné zejména po tělesné námaze.** Proto *bude ve sklenicích o obsahu 0.25 litru nabízeno například ve fitnesscentrech a na plovárnách.*	**The beer, the name of which should be decided in the course of this week is, according to his words, suitable especially after physical exercise.** Therefore, *it will be offered in 0.25 liter glasses for example in fitness centers and at swimming pools.*
	Edvard Beneš byl tématem natolik kontroverzním, že přivedl do varu i nejseriióznější historiky. **Není jim co závidět**: *Beneš patří mezi ty kultovní osobnosti, kterých si vážíme tím méně, čím více se o nich dovídáme.*	Edvard Beneš was a subject of so much controversy that he got inflamed even the most serious historians. **They are not to be envied**: *Beneš is one of those iconic figures we cherish the less, the more we learn about them.*
Pragmatic Reason – Result	**Při posuzování premiéra Klause bere veřejnost patrně v úvahu i jeho odpovědnost za činnost celého kabinetu, případně jednotlivých resortů.** *Dlouhodobé výsledky STEM* totiž *ukazují, že vždy, když klesala důvěra ve vládu, klesala i důvěra v premiéra.*	**When evaluating the Prime Minister Klaus, the public apparently takes into account his responsibility for the activities of the whole Cabinet or the individual departments.** *Long-time results of the STEM agency show* [totiž = roughly *as a matter of fact*] *that whenever the trust in government declined, the confidence in the prime minister dropped, too.*

Relation type	Czech example (PDiT 1.0)	English translation
Condition	**Posluchač musí přistoupit na pozici, že vše je dovoleno.** Potom *se pobaví a také pochopí, že drama znázorňuje ztrátu reálné komunikace.*	**The listener has to accept the position that everything is permitted.** Then *he enjoys himself and also understands that the drama symbolizes the loss of a real-life communication.*
Pragmatic Condition	Jestliže **chcete slyšet můj postoj k rozhodnutí poroty**, *je to neslýchaný projev neúcty k práci druhého.*	If **you want to hear my attitude towards the jury's decision**, *it is an outrageous sign of disrespect for the work of others.*
Purpose	*Odcizené věci si vojáci uložili do svých skříněk* s tím, že **si je odvezou do civilu.**	*The soldiers have stored the stolen things into their lockers* in order to **take them with them into civilian life.**
Explication	*Včerejší porada ministrů o státním rozpočtu na rok 1995 dopadla víc než dobře.* **Václav Klaus ani Ivan Kočárník** totiž **nenašli v Kramářově vile nikoho, kdo by se s nimi chtěl prát o ideu vyrovnaného rozpočtu.**	*Yesterday's meeting of the ministers concerning the state budget for 1995 ended better than well.* [Lit. As a matter of fact], **neither Václav Klaus nor Ivan Kočárník found anyone in the Kramář villa who would want to oppose them about the idea of a balanced budget.**
CONTRASTIVE RELATIONS		
Confrontation	*Stejně dobře vykročila i Radka Bobková, jež vyřadila domácí Poovou 3:6, 7:5, 7:6.* **Nedařilo se** naopak **Ludmile Richterové, jíž vystavila stop ve třech sadách 3:6, 6:2, 4:6 další domácí tenistka Werdelová.**	*Similarly well started also Radka Bobková who knocked out the domestic player Po 3:6, 7:5, 7:6.* On the contrary, **things did not go so well for Ludmila Richterová who was eliminated in three sets 3:6, 6:2, 4:6 by another domestic tennis player Werdel.**
Opposition	*Lidé chtějí platit jen to, co skutečně spotřebovali.* **Ještě dlouho tomu tak** ale **patrně nebude.**	*People want to pay for only what they really have consumed.* But **it apparently won't be this way for a long time yet.**
Pragmatic opposition	*Podle vedoucího výroby Miloše Příklopila má Seba rozpracovanou celou řadu zakázek.* **Zákazníci** však **vyvíjejí velký tlak na snižování cen tkanin.**	*According to the production manager Miloš Příklopil, the Seba company has a range of factory orders in process.* **The customers,** however, **exert great pressure on lowering of the prices of fabrics.**
Restrictive Opposition	*Každá krajina má svou krásu.* Jenom **ji musíte umět vidět.**	*Every landscape has its beauty.* Only **you must be able to see it.**
Concession	**Zdálo by se, že pirátské zboží zmizí z trhu.** Ale *po krátkém období paniky se překupníci a prodejci rychle vracejí k původní praxi.*	**It would seem that the pirate goods disappear from the market.** But *after a brief period of panic, the traffickers and the dealers quickly return to the original practice.*

Relation type	Czech example (PDiT 1.0)	English translation
Correction	[Opera Mozart] *Neprovozuje moderní hudební divadlo*, nýbrž **degraduje Mozartovu hudbu na pouhý kulisový doprovod k mnohdy samoúčelným jevištním skopičinkám.**	[Opera Mozart] *It does* not *perform modern musical theater*, **it** rather **degrades Mozart's music to a mere stage set accompaniment to often purposeless stage foolery.**
Gradation	*Letos se již zdálo, že počáteční nadšení místních radních pro tuto akci vychladlo.* **Organizátoři** dokonce **uvažovali o přemístění sympozia do Českých Budějovic.**	*This year it has seemed that the initial enthusiasm of local councilors for this action has faded.* **The organizers** even **considered relocating the symposium to České Budějovice.**
RELATIONS OF EXPANSION		
Conjunction	*Mövenpick provozuje několik desítek hotelů nejen v Evropě, ale i v Asii a Africe.* Kromě toho **je známý i jako obchodní a potravinářská firma.**	*Mövenpick operates dozens of hotels not only in Europe but also in Asia and Africa.* Apart from that, **it is known also as a business and food company.**
Instantiation	*Každá pověřená poradna spravuje agendu žadatelů o adopci v rámci větších územních celků.* Například **naše poradna v Kolíně působí ve dvanácti okresech středních Čech.**	*Each authorized advisory office administers an agenda of adoption applicants within larger areas.* For example, **our advisory office in Kolín operates in twelve districts of central Bohemia.**
Specification	*V souladu se západními vzory je možná i omezená preference soukromého pojištění před sociálním pojištěním.* Konkrétně, **pokud si výdělečně činná osoba zaplatí dostatečně vysoké soukromé pojištění, bude se moci ze sociálního pojištění „vyvléknout“.**	*In line with the western models, a limited preference of the private insurance over the social insurance is possible.* Specifically, **if an employed person pays a high enough private insurance, they can "wriggle out" of the social insurance.**
Equivalence	Dnes nebo zítra se v dolní komoře polského parlamentu - v Sejmu - očekává hlasování, které bude mít vážné politické důsledky, *ať už dopadne jakkoliv*, tj. **bude-li zákon odmítnut či přijat.**	Today or tomorrow the lower chamber of the Polish Parliament – the Sejm – expects voting that will have serious political consequences *whatever the outcome will be,* i.e. **whether the law will be rejected or accepted**.
Generalization	*Naše čtenářka, která by uzavřela životní pojištění na 20 let na pojistnou částku 100 tisíc s měsíčním pojistným 310 korun, by se mohla úrazově připojistit na dalších 100 tisíc za 32 korun měsíčně, zároveň by tím byla připojištěna i na úraz s trvalými následky na 200000.* **Ročně by** tedy **zaplatila na pojistném, včetně úrazového připojištění, 4104 korun.**	*Our reader, who would take out a life insurance for 20 years for an insured sum of 100,000 CZK with a monthly fee of 310 CZK could take out also an accident insurance for an additional 100,000 CZK for 32 crowns a month, at the same time she would be insured against an injury with permanent damage for 200,000 CZK.* Thus, **she would pay annually 4,104 crowns, including the accident insurance.**

Relation type	Czech example (PDiT 1.0)	English translation
Conjunctive alternative	[...schopní lidé se dnes již věnují pouze své profesi, neboť] *na amatérské působení mimo svou odbornost již nemají čas* nebo **se jim to prostě nevyplácí.**	[...talented people today are dedicated only to their profession, as] *they no longer have time for amateur activities outside their expertise* or **such activities just don't pay off.**
Disjunctive alternative	Proto je obzvlášť tristní poznání, že *vlády na krátící se termín blokace zákona o bankrotu zřejmě jednoduše zapomněly.* Nebo **mu nevěnovaly dostatečnou pozornost.**	It is particularly sad to realize that *the governments apparently simply forgot about the deadline for blocking the bankruptcy law.* Or **they just did not pay enough attention to it.**

Appendix 2:

Ambiguous connectives in the PDT 3.0 sorted according to the decreasing entropy (H) of their semantic categories – first 50 connectives

H (Entropy)	Connective	Semantic type	Occurrences	%
2.30	**když**	cond	240	41.74
		preced	116	20.17
		synchr	95	16.52
		spec	71	12.35
		reason	26	4.52
		conj	11	1.91
		conc	7	1.22
		explicat	5	0.87
		confr	3	0.52
		restr	1	0.17
2.26	**přitom**	conj	77	42.54
		conc	41	22.65
		opp	36	19.89
		confr	8	4.42
		grad	6	3.31
		restr	4	2.21
		spec	4	2.21
		synchr	3	1.66
		reason	1	0.55
		f_opp	1	0.55
2.12	**- (Dash)**	spec	108	49.54
		reason	39	17.89
		conj	37	16.97
		explicat	21	9.63
		equiv	3	1.38
		confr	2	0.92
		opp	2	0.92
		gener	2	0.92
		conjalt	1	0.46
		restr	1	0.46
		cond	1	0.46
		exempl	1	0.46

H (Entropy)	Connective	Semantic type	Occurrences	%
1.78	**aniž**	opp	21	41.18
		conj	20	39.22
		conc	7	13.73
		cond	1	1.96
		restr	1	1.96
		confr	1	1.96
1.69	**nicméně**	opp	41	57.75
		conc	16	22.54
		restr	7	9.86
		confr	5	7.04
		f_opp	2	2.82
1.64	**tedy**	reason	186	60.39
		gener	58	18.83
		equiv	50	16.23
		f_reason	4	1.30
		corr	3	0.97
		conj	2	0.65
		spec	2	0.65
		exempl	1	0.32
		opp	1	0.32
		conc	1	0.32
1.51	**a to**	spec	62	52.54
		conj	45	38.14
		grad	6	5.08
		conc	2	1.69
		reason	1	0.85
		restr	1	0.85
		opp	1	0.85
1.38	**: (Colon)**	spec	289	72.98
		reason	48	12.12
		explicat	32	8.08
		conj	18	4.55
		equiv	2	0.51
		exempl	2	0.51
		f_reason	2	0.51
		gener	1	0.25
		grad	1	0.25
		f_cond	1	0.25

H (Entropy)	**Connective**	**Semantic type**	**Occurrences**	**%**
1.36	**pak**	preced	190	64.19
		conj	77	26.01
		reason	20	6.76
		cond	7	2.36
		opp	1	0.34
		other	1	0.34
1.36	**tak**	reason	85	75.89
		gener	13	11.61
		equiv	6	5.36
		f_reason	2	1.79
		preced	1	0.89
		conj	1	0.89
		spec	1	0.89
		cond	1	0.89
		exempl	1	0.89
		confr	1	0.89
1.35	**či**	disjalt	54	62.79
		conjalt	24	27.91
		conj	6	6.98
		grad	1	1.16
		corr	1	1.16
1.26	**ovšem**	opp	228	77.82
		restr	25	8.53
		confr	17	5.80
		conj	9	3.07
		conc	8	2.73
		f_opp	3	1.02
		grad	2	0.68
		cond	1	0.34
1.19	**totiž**	reason	321	69.63
		explicat	112	24.30
		f_reason	18	3.90
		spec	7	1.52
		gener	1	0.22
		corr	1	0.22
		conj	1	0.22

H (Entropy)	**Connective**	**Semantic type**	**Occurrences**	**%**
1.17	**jenže**	opp	55	76.39
		restr	9	12.50
		confr	5	6.94
		conc	2	2.78
		f_opp	1	1.39
1.13	**avšak**	opp	49	80.33
		restr	4	6.56
		conc	3	4.92
		confr	3	4.92
		f_opp	1	1.64
		grad	1	1.64
1.12	**však**	opp	1240	81.20
		restr	90	5.89
		confr	85	5.57
		conc	57	3.73
		conj	28	1.83
		f_opp	18	1.18
		grad	7	0.46
		corr	2	0.13
1.02	**ale**	opp	1077	84.47
		conc	59	4.63
		restr	48	3.76
		confr	41	3.22
		corr	16	1.25
		conj	14	1.10
		f_opp	13	1.02
		cond	3	0.24
		grad	3	0.24
		reason	1	0.08
0.96	**navíc**	grad	112	61.54
		conj	70	38.46
0.91	**zároveň**	conj	63	67.02
		synchr	31	32.98

H (Entropy)	Connective	Semantic type	Occurrences	%
0.89	**nebo**	disjalt	138	72.25
		conjalt	52	27.23
		conj	1	0.52
0.8884	**potom**	preced	42	82.35
		cond	6	11.76
		conj	2	3.92
		equiv	1	1.96
0.78	**jestliže**	cond	73	87.95
		f_cond	4	4.82
		confr	3	3.61
		conc	1	1.20
		conj	1	1.20
		reason	1	1.20
0.77	**sice ale**	opp	148	88.10
		conc	8	4.76
		restr	5	2.98
		f_opp	3	1.79
		confr	3	1.79
		corr	1	0.60
0.72	**přičemž**	conj	76	88.37
		opp	4	4.65
		spec	4	4.65
		grad	1	1.16
		confr	1	1.16
0.68	**i když**	conc	160	89.89
		opp	6	3.37
		restr	6	3.37
		f_opp	3	1.69
		cond	2	1.12
		confr	1	0.56
0.65	**a pak**	preced	49	87.50
		conj	5	8.93
		cond	2	3.57

H (Entropy)	Connective	Semantic type	Occurrences	%
0.57	**zatímco**	confr	181	88.73
		synchr	21	10.29
		opp	1	0.49
		conj	1	0.49
0.56	**naopak**	confr	137	90.13
		opp	8	5.26
		corr	7	4.61
0.56	**i**	conj	76	91.57
		grad	4	4.82
		reason	1	1.20
		conc	1	1.20
		disjalt	1	1.20
0.53	**takže**	reason	137	91.95
		gener	7	4.70
		equiv	2	1.34
		f_reason	2	1.34
		explicat	1	0.67
0.40	**poté co**	preced	81	94.19
		explicat	3	3.49
		reason	1	1.16
		conc	1	1.16
0.39	**aby**	purp	290	94.77
		reason	8	2.61
		conj	5	1.63
		preced	1	0.33
		f_reason	1	0.33
		explicat	1	0.33
0.36	**dokonce**	grad	68	93.15
		conj	5	6.85
0.32	**a tak**	reason	135	95.74
		conj	3	2.13
		equiv	2	1.42
		gener	1	0.71

H (Entropy)	Connective	Semantic type	Occurrences	%
0.30	**což**	conj	182	96.30
		opp	3	1.59
		equiv	2	1.06
		spec	1	0.53
		reason	1	0.53
0.28	**neboť**	reason	213	96.38
		explicat	5	2.26
		other	1	0.45
		f_reason	1	0.45
		spec	1	0.45
0.22	**proto že**	reason	96	96.97
		explicat	2	2.02
		f_reason	1	1.01
0.20	**li**	cond	243	97.59
		f_cond	3	1.20
		confr	2	0.80
		purp	1	0.40
0.20	**přesto**	conc	96	96.97
		opp	3	3.03
0.20	**dále**	conj	114	97.44
		preced	2	1.71
		reason	1	0.85
0.16	**přestože**	conc	96	97.96
		confr	1	1.02
		opp	1	1.02
0.16	**protože**	reason	515	98.10
		explicat	7	1.33
		f_reason	2	0.38
		cond	1	0.19

H (Entropy)	Connective	Semantic type	Occurrences	%
0.16	**proto**	reason	373	98.16
		f_reason	4	1.05
		explicat	2	0.53
		equiv	1	0.26
0.14	**kdyby**	cond	114	98.28
		corr	1	0.86
		conc	1	0.86
0.14	**a také**	conj	50	98.04
		cond	1	1.96
0.13	**a**	conj	5746	98.73
		reason	21	0.36
		opp	16	0.27
		confr	13	0.22
		spec	8	0.14
		equiv	3	0.05
		preced	3	0.05
		synchr	3	0.05
		cond	2	0.03
		conc	2	0.03
		grad	1	0.02
		disjalt	1	0.02
		restr	1	0.02
0.11	**pokud**	cond	399	98.76
		f_cond	4	0.99
		grad	1	0.25
0.08	**například**	exempl	96	98.97
		reason	1	1.03
0.08	**#neg ale**	corr	97	98.98
		opp	1	1.02
0.08	**rovněž**	conj	106	99.07
		opp	1	0.93