STUDIES IN COMPUTATIONAL
AND THEORETICAL LINGUISTICS

DISCOURSE AND COHERENCE

From the Sentence Structure to Relations in Text

Šárka Zikánová, Eva Hajičová, Barbora Hladká, Pavlína Jínová, Jiří Mírovský, Anna Nedoluzhko, Lucie Poláková, Kateřina Rysová, Magdaléna Rysová, Jan Václ

ÚFAL
ÚSTAV FORMÁLNÍ
A APLIKOVANÉ LINGVISTIKY

DISCOURSE AND COHERENCE

From the Sentence Structure to Relations in Text

Šárka Zikánová, Eva Hajičová, Barbora Hladká, Pavlína Jínová,
Jiří Mírovský, Anna Nedoluzhko, Lucie Poláková,
Kateřina Rysová, Magdaléna Rysová, Jan Václ

ÚFAL
ÚSTAV FORMÁLNÍ
A APLIKOVANÉ LINGVISTIKY

STUDIES IN COMPUTATIONAL AND THEORETICAL LINGUISTICS

Šárka Zikánová, Eva Hajičová, Barbora Hladká, Pavlína Jínová, Jiří Mírovský, Anna Nedoluzhko, Lucie Poláková, Kateřina Rysová, Magdaléna Rysová, Jan Václ

DISCOURSE AND COHERENCE
From the Sentence Structure to Relations in Text

Published by the Institute of Formal and Applied Linguistics
as the 14th publication in the series
Studies in Computational and Theoretical Linguistics.

The authors gratefully acknowledge the support of the grants GAP406/12/0658 (Coreference, discourse relations and information structure in a contrastive perspective), LM2010013 (LINDAT-CLARIN – Establishing and operating the Czech node of pan-European infrastructure for research), LH14011 (Multilingual Corpus Annotation as a Support for Language Technologies), P46 (PRVOUK – Programs of development of scientific areas at the Charles University in Prague: Informatics) and of the institutional funds of the Charles University in Prague.

Printed by Printo, spol. s r. o.

ISBN 978-80-904571-8-8

Contents

Preface **1**

1 Introduction **3**

General Background **11**

2 Discourse Relations **13**

2.1 Discourse Relations 13

2.2 The Penn Discourse Treebank 14

2.3 Discourse Connectives 15

2.4 Discourse Arguments 17

2.4.1 Notation of the arguments 18

2.5 Semantic Types of Discourse Relations 18

2.6 Annotation Process 22

2.6.1 Theoretical starting points 22

2.6.2 Inspiration from the PDTB approach 23

2.6.3 Annotating on top of syntactic trees 23

2.6.4 Two-phase annotation 24

2.7 Other Annotated Phenomena 26

2.7.1 List structures 26

2.7.2 Discourse special: headings, captions, metatexts 27

2.7.3 Genre annotation 27

2.8 Summary 27

3 Coreference **29**

3.1 Basic Terms 29

3.2 Related Work 32

3.3 Grammatical and Textual Coreference 33
3.3.1 Grammatical coreference 34
3.3.2 Textual coreference 36
3.4 Coreference of Nominal Groups with Different Referential Potential 39
3.5 Coreference Annotation in the PDT 43
3.5.1 Scope of annotated expressions 44
3.5.2 Embedded nominal groups 44
3.5.3 Syntactic zeros 45
3.5.4 Non-referring expressions 48
3.5.5 Coordinative constructions 48
3.5.6 Coreference with specific and generic nominal groups 49
3.5.7 Discourse deixis 50
3.5.8 Prepositional phrases 50
3.6 Summary 50

4 Bridging Relations 51

4.1 Typology of Bridging Relations 51
4.2 Annotation of Bridging Relations in Corpora 54
4.3 Annotation of Bridging Relations in the PDT 54
4.3.1 Meronymical relation between a part and a whole 55
4.3.2 The relation between a set and its subsets/elements 56
4.3.3 The relation between an entity and its singular function 58
4.3.4 The relation between coherence-relevant discourse opposites 58
4.3.5 Non-coreferential explicit anaphoric relation 59
4.3.6 Further underspecified bridging relations 60
4.4 Discussion and Summary 61

5 Topic–Focus Articulation 63

5.1 What Is Topic–Focus Articulation 63
5.2 The Importance of Topic–Focus Articulation 64
5.3 The Theoretical Basis 64
5.4 Basic Terms of Topic–Focus Articulation 67
5.4.1 Context and contextual boundness 67
5.4.2 Communicative dynamism 69
5.4.3 Topic and focus 71

5.5 Detection of Topic and Focus 73
5.5.1 Question test 74
5.5.2 Test with negation 75
5.6 Representation of TFA in the Prague Dependency Treebank 76
5.6.1 Annotation of contextual boundness in the PDT 76
5.6.2 Annotation of communicative dynamism in the PDT 78
5.7 Summary 80

Data **81**

6 Prague Dependency Treebank **83**
6.1 Layers of Annotation 84
6.2 Discourse Coherence Phenomena 86

7 Inter-Annotator Agreement **89**
7.1 Within a Single Sentence 90
7.2 Crossing the Sentence Boundary 92
7.3 At the Document Level 94
7.4 Summary 94

8 Searching in the PDT **97**
8.1 Basics of the PML-TQ Language 98
8.1.1 Node selection 98
8.1.2 Relations between nodes 99
8.1.3 Negative query 101
8.1.4 Crossing the layers of annotation 102
8.2 Discourse Coherence Phenomena and the PML-TQ 103
8.2.1 Non-dependency relations 103
8.2.2 Topic–focus articulation and anaphora 105
8.2.3 Output filters 107
8.2.4 Output filters in discourse 108
8.3 Hands on the Data 111
8.3.1 Data to download 111
8.3.2 Data for searching 114

Case Studies 115

9 Relation of Discourse Analysis to Syntax 117

9.1 Features of Syntactic Analysis Used for Discourse-Level Analysis 118

9.1.1 Syntactico-semantic labels for relations between clauses 121

9.1.2 Scope of discourse arguments 132

9.1.3 Connectives 136

9.2 Discourse Structure from the Syntactic Point of View 139

9.2.1 Discourse relations realized intra-sententially and inter-sententially . 140

9.2.2 Discourse relations in subordinate versus coordinate structures . . . 143

9.3 Summary 146

10 Morphosyntactic Characteristics of Czech Connectives 149

10.1 General Characteristics 149

10.1.1 Part-of-speech classification 150

10.1.2 Form and inflection 151

10.1.3 Origin 153

10.1.4 Placement in the sentence and in the argument 155

10.1.5 Subordinate, coordinate and inter-sentential connectives 156

10.2 Characteristics of Most Frequent Connectives 157

10.2.1 Frequency 157

10.2.2 Part-of-speech characteristics 159

10.2.3 Intra- and inter-sentential use of connectives 160

10.2.4 Degree of connectivity 161

10.3 Summary 163

11 Multiword Discourse Phrases 165

11.1 A Scale of Explicitness and Implicitness of Discourse Relations 165

11.2 Terminology 166

11.3 Current Annotation of Secondary Connectives in the PDT 167

11.4 Lexico-Syntactic Characteristics of English AltLexes in the PDTB 167

11.5 Syntactic Characteristics of Czech Secondary Connectives 168

11.5.1 Secondary connectives realized by verbal phrases 171

11.5.2 Prepositional phrases 171

11.5.3 Secondary connectives realized by (semi-)clauses 172

11.6 Lexical Characteristics of Secondary Connectives in Czech 173
11.7 Semantic Characteristics of Secondary Connectives in Czech 174
11.8 Summary 176

12 Exploration of Weak Coherence and Coherence Disruptions 179
12.1 Terminology, Data and Workflow 180
12.1.1 Unsignaled relations in the RST Discourse Treebank 180
12.1.2 Treatment of no relation in the PDTB 181
12.1.3 Treatment of no relation in the PDT 182
12.2 Results and Discussion 185
12.2.1 New types of relations 187
12.2.2 Reader’s expectation as coherence factor 190
12.3 Summary 195

13 Contextually Bound Expressions without a Coreference Link 197
13.1 Data 199
13.2 Reasons of Contextual Boundness without Anaphoric Links 200
13.2.1 Deduction of contextual boundness from previous context 202
13.2.2 Extralinguistic reasons for contextual boundness 204
13.2.3 “Scene setting” circumstances 206
13.2.4 Contextually bound expressions representing measures 207
13.2.5 Technical reasons 209
13.3 Discussion 211
13.4 Summary 213

14 Tracing Salience 215
14.1 Motivation 215
14.2 Related Work 217
14.3 Related Linguistic Phenomena annotated in the PDT 219
14.4 The Salience Algorithm 223
14.5 Learning Salience 229

15 Summary 233

CONTENTS

List of Abbreviations 239

Bibliography 241

Sources 257

Subject Index 259

Name Index 263

Motto:

Vůbec se mi zdá, že nejlepší myšlenka je ta, která ponechává vždy určitou skulinu pro možnost, že všechno je zároveň úplně jinak.

The best possible idea, I believe, is one that always leaves room for the possibility that things are, at the same time, utterly different.

— Václav Havel

Preface

In this monograph we present the results of our research on the interplay of *intra-sentential* relations such as deep syntactic relations and information structure of the sentence and the *inter-sentential* relations such as discourse relations and coreferential and other associative links. The book is a collective work and all the authors share the responsibility for revising and editing all chapters, and ultimately for the content of the chapters. On the other hand, each chapter has different primary authors. The primary authors are as follows: Eva Hajičová (Chapter 1), Lucie Poláková (Chapter 2 and co-author of Chapter 10), Anna Nedoluzhko (Chapters 3, 4 and Chapter 13), Kateřina Rysová (Chapter 5), Jiří Mírovský (Chapters 6 through 8), Pavlína Jínová (Chapter 9 and co-author of Chapter 10), Magdaléna Rysová (Chapter 11), Šárka Zikánová (Chapter 12) and Barbora Hladká and Jan Václ (Chapter 14). The research presented in this monograph was carried out on a large corpus of Czech language data annotated by the authors themselves but also with the help of a team of student annotators; their involvement in the project and their highly time-consuming work was extremely valuable and our sincere thanks go to them as well.

The authors also highly appreciate the final text revisions made by Barbora Štěpánková and the language revision carried out by Masha Volynsky. We are indebted to the two reviewers Ekaterina Lapshinova-Koltunski and Maciej Ogrodniczuk for their detailed and insightful comments, which were most valuable for the final wording of the text. We take full responsibility, of course, for its remaining shortcomings. We thank to Eduard Bejček for technical assistance and Petra Hoffmanová for comments and suggestions on writing the bibliography.

Last but not least, our gratitude goes to our colleagues in the Institute of Formal and Applied Linguistics at the Faculty of Mathematics and Physics of Charles University in Prague, for their support and for providing a most friendly and stimulating working atmosphere. Without this moral, as well as intellectual, support our work could not have reached its goal.

The authors gratefully acknowledge the support provided by the following grants: GAP406/12/0658 (Coreference, discourse relations and information structure in a contrastive perspective), LM2010013 (LINDAT–CLARIN – Establishing and operating the Czech node of pan-European infrastructure for research), LH14011 (Multilingual Corpus Annotation as a Support for Language Technologies), and P46 (PRVOUK – Programs of development of scientific areas at the Charles University: Informatics).

1
Introduction

Since the last decades of the twentieth century, a strong and influential tendency in linguistic studies has developed, that has moved away from the traditional emphasis on sentence syntax and semantics towards research focusing on text and discourse, or, at least, widened the range of linguistic investigations from matters of linguistic competence to regularities in the use of language or "communicative competence" (Sgall, Hajičová and Panevová, 1986). This shift raised a number of research questions: What is the nature of text? Are there general rules for the structure of text? If so, what is the mechanism that enables competent speakers to use the language they have internalized in order to communicate with other speakers? What is the relation of the evolving text linguistics to the traditional fields such as stylistics and rhetoric? Is it possible in the study and description of text structure to employ methods of formal logic, which have already been applied for an account of various phenomena not only within syntax and semantics, but also pragmatics?

The range of literature devoted to the above issues as well as to different aspects of the structure of text is very broad (see the references throughout this monograph) and thus one may ask why enlarge it with another book on text or discourse.[1] When studying the relevant literature we have noticed one prevailing feature of the available resources: Authors mostly concentrate either on the general issues as listed above or on one aspect of the analysis of text or discourse structure. Our view may be called holistic – we follow and analyze different aspects of discourse structure with regard to their interplay in the constitution of an integrated whole, a coherent (segment of) text.[2]

Coherence and *cohesion* (cf. de Beaugrande and Dressler, 1981) are the most important constitutive features of text, or, in other words, of textuality. These two terms are often used as synonyms; if differentiated, the former refers to the conceptual and semantic dimension of text and the integration of individual conceptual segments into an integrated whole, while the latter refers to the expressive means of the build-up of such a whole (cf. Hoffmannová, 1993).

There are many factors that are involved in making discourse an integrated whole. Halliday and Hasan (1976) in their classical and most detailed analysis of cohesion

[1] In this chapter we use the terms *discourse* and *text* as rough synonyms that came into existence for more or less historical or geographical reasons.

[2] In a certain respect, we follow a strategy similar to that of Grosz and Sidner (1986) who discuss the mutual relationships of three structures, namely the linguistic structure, the attentional state and intentional structure.

and coherence distinguish five such aspects that together organize a text as "a neatly woven texture": conjunctions, reference, substitution, ellipsis and lexical cohesion. There are, of course, many other points of view that can be applied in discourse analysis, be it the intentional structure of a discourse, the discourse communicative functions, speech act analysis, the so-called pragmatic discourse relations, the subjectivity of discourse, inferences that can be drawn from a discourse segment, etc., to name just a few. In our analysis we concentrate on the following factors we believe to be crucial and to play an integrating role, though we are aware that the list of aspects we focus our attention on is far from being exhaustive:

(i) Since the building stone of discourse is a sentence, we study in which respects the sentence structure itself contributes to discourse structure; we base our analysis on the *deep syntactic structure* of the sentence.[3] We pay special attention to the *information structure of the sentence* (its *topic–focus articulation*) which is supposed to be an integral part of the deep syntactic structure. We also apply the information structure analysis together with the analysis of coreference links in order to follow the development of discourse in terms of the *salience* of the elements of the stock of knowledge assumed by the speaker to be shared by him and the hearer.

(ii) One distinctive feature of our methodology is the fact that we build the discourse relations on top of the *deep (underlying) dependency structure* of sentences rather than on the raw text, which makes it possible to follow in which respects a representation of this structure can help us to identify discourse relations and their scope.

(iii) Moving from the constituting elements of the discourse to the relations that combine these elements into larger wholes, or, more specifically, that exist between elementary parts of discourse, we analyze and classify the so-called *discourse relations* and look for the *linguistic means* identifying them; these means include connectives or some alternative complex expressions. We do not exclude the so-called *implicit* relations, i.e. those that are not expressed explicitly.

(iv) An invaluable contribution to the connectivity of discourse is played by the connective threads carried out via *coreference links* and other *associative relations*.

Before we devote our attention to these factors in greater detail, let us illustrate the interplay which forms the background of our consideration on a piece of a continuous text. The text is a considerably shortened extract (p. 251 ff.) of Josef Škvorecký's book *Dvorak in Love. A light-hearted dream* (translated from the Czech original *Scherzo capriccioso* by Paul Wilson, published by Lester & Orpen Dennys Limited, Toronto in 1986). The point of the extract is to fabulate a story about the world-famous Czech composer Antonín Dvořák, namely how the idea of the composition of the opera

[3] See below concerning the notion of deep (tectogrammatical) structure in our approach to a multilevel description of language.

Rusalka ("a water nymph") came to him. The story talks about how two youngsters, Dvořák's daughter Magda and her boy-friend Kovarik, went out for a walk (probably without her father's knowledge) along the Turkey river.[4]

> (1) *Across the river Magda and Kovarik could now see a fire with two figures beside it.* (2) *When they moved closer,* (3) *they could make out two white horses against the background of the dark bushes.* (4) *Then he recognized them.* (5) *The pale blue buggy.* (6) *Two hours ago, the beauty from Chicago had sat on the seat* (7) *while the black man in livery had gone into Kapino's for beer.* (8) *They stopped* (9) *and looked across the river.* (10) *The young lady in the white dress was biting into a chicken leg.* (11) *He looked at Magda.* (12) *The child's eyes, wide in amazement, stared across the river at this fairy-tale banquet.* (13) *He looked at the straw hat.* (14) *Yes, beside it in the grass a pair of white shoes had been casually tossed* (15) *and beside them lay a crumpled white pile.* (16) *The beauty stood up* (17) *and threw the half-eaten leg into the fire.* (18) *She stretched.* (19) *She said something to the man.* (20) *She lifted up her skirt* (21) *and, stepping gingerly through the grass,* (22) *she began walking upstream.* (23) *Her head became a cooly glowing torch.* (24) *Intoxitated, Kovarik stepped forward* (25) *and silently followed the beautiful phantom's pilgrimage.* (26) *From downstream they could hear a banjo playing.* (27) *A pleasant baritone voice sang: "…".* (28) *The girl let her hands drop.* (29) *Cautiously, she stepped into the water.* (30) *On their side of the river, something creaked.* (31) *Looking towards the sound, he could barely distinguish the outline of a small rowboat* (32) *and, in it, someone's dark silhouette.* (33) *The moonlight fell on the head, the white whiskers, the hair in disarray.* (34) *The Master!* (35) *He looked quickly across the stream* (36) *and saw the Rusalka up to her waist in the water.* (37) *"Borne like a vapour…"* (38) *The Rusalka was slowly lowering herself into the water.* (39) *Finally, all that remained on the water was a burning waterlily.* (40) *Suddenly the child saw too* (41) *and shrieked,* (42) *"Papa!"* (43) *The Master looked around* (44) *and then saw.* (Škvorecký, 1986)

The influence of the *information structure* on the choice of referring expressions is reflected in sentence (6): The use of the definite noun group *the beauty from Chicago* in the topic part of the sentence is conditioned by the fact that the referent of this noun is known from the previous context (this contextual knowledge is indicated by sentence (4)), otherwise the referent should be introduced in the focus part of the sentence. The same is true about the referent of the definite noun group *the black man in livery* in sentence (7). Sentence (5) is a topicless sentence, the noun group *the pale blue buggy* being its focus. However, the use of the definite article indicates that the sentence can be understood as standing in an implicit specification relation to the previous

[4] We number the sentences or their parts in order to make it easier to refer to them in the following analysis but we do not separate them on extra lines to make the flow of the discourse uninterrupted.

sentence and relating the buggy (by means of the use of the pronoun *them*) to the two figures and the two horses. From the point of view of the development of the *salience* of the individual elements of the stock of knowledge, it can be observed that some of the referents keep their position on the top of the stock for the whole of the story – this concerns both of the youngsters – while some emerge at some moment and fade away (the black man), some enter the scene at a later moment and stay (the lady) and some appear suddenly at a later stage and stay (the Master). These movements and changes in activation are reflected in the segmentation of the discourse and in the identification of the topics of discourse.

The contribution of the *sentence structure* is manifested e.g. in the relation of (2) and (3) which is as a matter of fact an intra-sentential relation of a dependent temporal clause (2) to its governor (3); the same holds true about the relation between (6) and (7), the latter being a temporal clause depending on the governor (6). Both of these relations are captured in the dependency-based deep syntactic structure of the complex sentences.

Discourse relations in the sense indicated above in points (ii) and (iii) are manifold, complex and often difficult to classify, and they are rendered by a number of linguistic means. The type of the relation can be deduced from some explicit one-word connective: e.g. *then* in (4) and (44), *finally* in (39), *and* in (15), (17), (32), (36) and (41) (with different implications of the type of relation: simultaneity in (15) and (32), posteriority in (17), (36) and (41)). The absence of an explicit connective does not necessarily mean an absence of a discourse relation: If we look at the sequence of (17) through (22), these sentences are linked as if there were a conjunction of coordination between each of the two clauses, partly interpreted as a simultaneity, partly as a succession. It is an open question how the English *-ing* form is to be interpreted: Does it function as an explicit discourse relation marker? Or is the relation between the clauses in which one includes a verb in the *-ing* form to be considered as an implicit discourse relation? In addition to connectors specified as one-word connectives there are other means of expressing discourse relations, namely multiword discourse phrases. There are no such connectives in the above extract but it can be easily imagined that the simple connective *then* in (4) is replaced by a complex expression *at that moment* or that the sentence (16), without an explicit relation marker, can be reformulated as *After a while, the beauty stood up* with an addition of an explicit expression rendering a temporal relation to the preceding sentence (15).

Coreference and associative relations seem to be the strongest cohesive means, though in many cases accompanied by ambiguity (or vagueness) of reference. This fact is reflected throughout the whole example text. Who is *they* in (2)? The ambiguity is resolved only by the following sentence because it could be only Magda and Kovarik who can be interested in making out what is happening on the other side of the river. A similar uncertainty concerning the reference concerns the pronoun *them* in (4): Does the pronoun refer to the figures or to the horses? Or to both? Similarly for *they* in (8): Who stopped? Magda and Kovarik or the lady and the black man? Who is *the girl* in (28)? Probably Magda, but the noun can also refer to the woman on the other side of the river. Actually, it is not before sentence (36) that we can establish for sure that the reference to the girl (and subsequent references by the feminine pronoun) coreferred to the lady on the other side of the river. Or was it the lifting of the skirt referred to in (20) that indicated who stepped into the water? A real puzzle is the reference by the pronoun *he* in (35). Only after reaching (44), we can decide that the pronoun in (35) referred to Kovarik. So far, we discussed the relation of coreference, i.e. the reference to the same referent (object). However, several associative relations appear in the text that contribute to its coherence: Thus the expressions *the straw hat*, *a pair of white shoes*, *a crumpled white pile*, *head*, *the phantom*, *a cooling torch*, *waterlilly* are associated with the lady, in a similar vein as the expressions *a banjo playing* and *a pleasant baritone voice* are in association to the black man, or *the seat* is related to the buggy and *the half-eaten leg* to the chicken. Such associative relations may be of different degrees of closeness and may be classified as different types.

We have used this illustrative example to indicate the richness and at the same time the interrelatedness of the three aspects we follow in the make-up of a coherent piece of discourse. In the chapters that follow, we analyze these aspects in detail using the material of the Prague Dependency Treebank, an annotated electronically available corpus of texts.[5]

The annotation scheme of the PDT is based on a solid, well-developed theory of an (integrated) language description, the so-called Functional Generative Description (FGD, see e.g. Sgall, 1967a; Sgall et al., 1969; Sgall, Hajičová and Panevová, 1986). The principles of the FGD were formulated as a follow-up to the functional approach of the Prague School and in adherence with the strict linguistic methodological requirements introduced by N. Chomsky. The FGD framework has the form of a generative description that is conceived of as a multi-level system proceeding from linguistic function (meaning) to linguistic form (expression), i.e. from the generation of a deep syntactico-semantic representation of the sentence through the surface syntactic, morphemic and phonemic levels down to the phonetic shape of the sentence. From the point of view of formal grammar, both syntactic levels are based on the relations of dependency rather than constituency.

The main focus is placed on the deep syntactic level, called *tectogrammatical* (the term borrowed from Putnam's seminal paper on phenogrammatics and tectogrammatics; Putnam, 1961). On this level, the representation of the sentence has the form of a dependency tree, with the predicate of the main clause as its root; the edges of the tree represent the dependency relations between the governor and its dependents. Only the autosemantic (lexical) elements of the sentence attain the status of legitimate nodes in the tectogrammatical representation; functional words such as prepositions, auxiliary verbs and subordinate conjunctions are not represented by separate nodes

[5] Each example taken from the PDT is marked accordingly; examples taken from other sources are also easily identifiable. If there is no source cited, the examples are our own.

and their contribution to the meaning of the sentence is captured by the complex labels of the legitimate nodes.

An important role in the derivation of sentences is played by the information on the *valency properties* of the governing nodes, which is included in the lexical entries: the valency values are encoded by the so-called *functors*, which are classified into *arguments* and *adjuncts*. We assume that each lexical entry in the lexicon is assigned a valency frame including all the obligatory and optional arguments appurtenant for the given entry; the frame also includes those adjuncts that are obligatory with the given entry; in accordance with the frame, the dependents of the given sentence element are established in the deep representation of the sentence and assigned an appropriate functor as a part of their complex label.

The representation of the sentence on the tectogrammatical level also captures the information structure of the sentence (its topic–focus articulation) by means of specifying individual nodes of the tree as contextually bound or non-bound and by the left-to-right order of the nodes. Coordination and apposition are not considered to be a dependency relation as they cannot be captured by the usual binary directional dependency relation. Coordinated sentence elements (or elements of an apposition) introduce a non-dependency, ″horizontal″ structure, possibly n-ary and/or nested, but still undirectional, where all elements have (in the standard dependency sense) a common governor (the only exception is formed by coordinated main predicates which naturally have no common governor). The coordinated (or appended) elements can also have common dependent(s). All the dependency relations expressed in a sentence with coordination(s) and/or apposition(s) can be extracted by ″multiplying″ the common dependency relations concerned. However, up to now, these relations have no direct counterparts in the FGD framework.

The *Prague Dependency Treebank* (see Chapter 6 below for a brief description and for references) consists of continuous Czech texts mostly written in journalistic style (taken from the Czech National Corpus)[6] analyzed on three levels of annotation (morphological, surface syntactic shape and underlying syntactic structure). At present (PDT 3.0 version), the total number of documents annotated on all the three levels is 3,165, amounting to 49,431 sentences and 833,193 (occurrences of) nodes. For the purpose of our analysis, a crucial role is played by the tectogrammatical layer capturing the underlying (“deep”) syntactic relations: The dependency structure of a sentence on this layer is a tree consisting of nodes only for autonomous meaningful units (as was already said, function words such as prepositions, subordinate conjunctions, auxiliary verbs etc. are not included as separate nodes in the structure and their contribution to the meaning of the sentence is captured by complex symbols of the autonomous units). Every node of the tectogrammatical representation is assigned a label consisting of: the lexical value of the word, its *(morphological) grammatemes*

[6] These texts became later part of corpora SYN2000 and SYN2006pub in the Czech National Corpus, available from https://www.korpus.cz.

(i.e. the values of morphological categories such as Feminine, Plural, Preterite etc.), its *functors* (such as *Actor*, *Patient*, *Addressee*, *Origin*, *Effect* and different kinds of circumstantials, with a more subtle differentiation of syntactic relations by means of *subfunctors*, e.g. *in*, *at*, *on*, *under*), and the *topic–focus articulation* (information structure, TFA) attribute containing the values for contextual boundness, on the basis of which the topic and the focus of the sentence can be determined. Pronominal and grammatical coreference is also annotated. It should be noted that the tectogrammatical representations may contain nodes not present in the morphemic form of the sentence in the case of surface deletions. In the process of further development of the PDT, additional information has been added to the original one, such as the annotation of multiword expressions, of basic relations of textual coreference and relations of association and of discourse relations.

In spite of the fact that the language material on which the analyses proposed in this monograph are carried out is a corpus of Czech, we hope that the basic conclusions we have reached have a more general validity. It is undisputable, however, that the typological properties of Czech language have to be taken into account. First, and most importantly, Czech is a language with rich inflection both in the nominal and verbal categories: with nouns, 7 cases, 2 numbers (with a relic of dual as a third member of the category) and 4 grammatical genders (masculine animate and inanimate, feminine and neuter) can be distinguished; with verbs, apart from person, number, tense, voice, and mode, a rather complex category of aspect (such as perfective and imperfective) is a prominent phenomenon. Together with rich inflection, we can also speak about the flexibility of Czech word order. In contrast to the grammatically fixed English, the word order in Czech is usually referred to as free; however, it is evident that it is not truly free but mostly guided by the information structure of the sentences. Another feature of Czech that is relevant for our analysis is the lack of determiners expressing definiteness and indefiniteness. Czech uses a variety of strategies instead, such as demonstrative and other kinds of pronouns, explicit phrases or even word order. Also connected with the inflectional character of Czech is its pro-dropness character: Personal pronouns of 1st and 2nd pers. singular and plural in the subject position can be in principle elided and their presence in that position is more or less marked. In contrast e.g. to English, Czech is also characterized by the possibility of "null" subjects.

The structure of the present monograph corresponds to our starting position and research methodology: In the part General Background we present an analysis of the aspects of discourse briefly outlined above (discourse relations in Chapter 2, coreference in Chapter 3, bridging relations in Chapter 4 and sentence information structure in Chapter 5). The theoretical considerations are followed, in the part Data, by a more detailed description of the language data used for our analysis (Chapter 6) and a statistical evaluation of the inter-annotator agreement that documents the different degrees of difficulty of the annotation tasks, and, consequently, the different degrees of complexity of the task of discourse analysis (Chapter 7). How the data can be searched is briefly discussed in Chapter 8. In the part Case Studies, we focus on some particular issues that emerged during our research and that deserve, in our opinion, a more detailed discussion: Among them are the relations between the syntactic structure of the sentence and discourse relations (Chapter 9), morphosyntactic characteristics of connective expressions in Czech (Chapter 10) and multiword connective phrases expressing discourse relations (Chapter 11), and cases where apparently there is no coreference link leading from a contextually bound element of the sentence (Chapter 13). The places with a weak coherence are discussed in Chapter 12 and a proposal on how to combine several aspects of discourse to trace salience of elements of the stock of shared knowledge is presented in Chapter 14.

General Background

2

Discourse Relations

One aspect of discourse coherence that has been at the center of interest to the discourse-oriented research community in the recent years are *discourse relations*. In this chapter, we describe the general features of this phenomenon and then focus on a more specific characterization motivated by the annotation-based decisions regarding their representation in the Prague Dependency Treebank. The chapter summarizes the research on the subject spanning across several years, and as a result it is largely based on previously published work: work-in-progress reports (Mladová, Zikánová and Hajičová, 2008; Jínová, Mírovský and Poláková, 2012a etc.), annotation guidelines (Poláková et al., 2012a), treebank introducing articles (Poláková et al., 2013; Zikánová et al., 2015) and a dissertation thesis (Poláková, 2015).

The term *discourse relations* has two interpretations. The broader one refers to all relations in discourse, including e.g. coreference and bridging relations, thematic structure etc. Throughout this book, and in accordance with the Penn Discourse Treebank terminology (Miltsakaki et al., 2004), we use this term in a narrower sense: The term *discourse relations* refers only to coherence relations that express a semantic connection between two discourse segments. The terminology used in the different approaches to describe these relations varies significantly. They may be called: *coherence relations* (e.g. Hobbs, 1979; Kehler, 2002), *rhetorical relations* (Mann and Thompson, 1988, Asher and Lascarides, 2003), *conjunctive relations* (Martin, 1992), *informational coherence relations* (Wolf and Gibson, 2005) and so on.

For the broader sense, to avoid ambiguity, we prefer to use the terms *coherence relations* or *relations in discourse*.

2.1 Discourse Relations

In this monograph, discourse relations are understood as semantic relations that connect two discourse units (segments of text expressing mostly individual events, states, situations). Discourse relations are often signaled by an explicit discourse-structuring device, like conjunctions, sentence adverbs etc. Example 1 repeats the first three sentences of the introductory text from J. Škvorecký (1986), and demonstrates the different realizations of discourse relations.[7]

[7] Depending on the definition of a discourse unit (henceforth *discourse argument*), there may be different analyses. For our purposes, the "smallest" discourse argument is represented by a simple clause with one predication. Hence, there are four discourse arguments in Example 1. More details on the delimitation and nature of discourse arguments are given in Section 2.4.

(1) (a) *Across the river Magda and Kovarik could now see a fire with two figures beside it.*
(b) *When they moved closer,*
(c) *they could make out two white horses against the background of the dark bushes.*
(d) *Then he recognized them.* (Škvorecký, 1986)

In Example 1, the discourse relation of the second sentence (arguments b and c) to the third sentence (argument d) is inter-sentential and it is explicitly signaled by the connective *then*. It expresses temporal succession of the events described by the arguments. Further, the first and the second sentence of the extract are connected mainly by means of a coreference link (*Magda and Kovarik – they*). The discourse relation between these two arguments is semantically not strongly perceived, yet it exists. It can be treated as a loose continuation, conjunction or succession of events with no explicit connective present.[8]

Finally, as follows from the delimitation of a discourse argument as a single clause, discourse relations can be intra-sentential, e.g. they may hold within individual sentences. Within the second sentence, the dependent clause (argument b) relates to its governing clause (argument c) also with the discourse relation of temporal asynchrony (succession of events). Note that the expression *and* in the first sentence does not function as a discourse connective in the given context. As a mere conjunction of entities it plays no role in the analysis of discourse relations.

2.2 The Penn Discourse Treebank

The analysis outlined above stems from two main sources of inspirations: some of its features are based on the Prague Functional Generative Description (FGD), in particular on the *tectogrammatical representation* of a sentence and its syntactico-semantic labels (called *functors*, cf. Chapter 9), but, more importantly, it is to a large extent inspired by the description of discourse relations in the Penn Discourse Treebank 2.0 (PDTB).

The PDTB annotation project is a lexically based model of discourse developed at the University of Pennsylvania (Miltsakaki et al., 2004; Prasad et al., 2008). The analysis of discourse relations in the PDTB consists primarily in finding and analyzing lexical cues 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 (Webber, Knott and Joshi, 2001). In the annotation scheme, discourse connectives include coordinating conjunctions, subordinating conjunctions and discourse adverbs.

Apart from connectives, the two discourse arguments of a discourse relation and the *semantic type* (sense) of a discourse relation were annotated. Discourse arguments

[8] According to some newer studies (e.g. Taboada and Das, 2013), the use of demonstrative pronouns and their referring potential can be interpreted as a kind of discourse-structuring device, although not as an actual discourse connective.

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 a connective is marked as Argument 2, the other as Argument 1, disregarding its location. For ascribing semantic categories to individual discourse connective occurrences, a set of 30 semantic labels was developed and organized in a three-level hierarchy (Prasad et al., 2007), with four semantic categories at the most general level (class level), further 16 categories on the second level (type level) and some of the types are further subcategorized into subtypes on the third, most fine-grained level.

In 2004, the first version of the Penn Discourse Treebank was released (Miltsakaki et al., 2004). The second release four years later includes manual annotation of approx. 49 thousand English sentences from the journalistic domain (PDTB 2.0; Prasad et al., 2008) for a given set of approx. 100 types of discourse connectives, their arguments and senses. A third version of the PDTB is a work in progress concentrating on annotation of intra-sentential discourse phenomena such as free adjuncts (Prasad et al., in prep.). In the second version so far, apart from explicit connectives, other phenomena have been annotated, mainly *implicit relations* (discourse relations that are not signaled by explicit connectives and must be inferred by the reader) and *attribution* (ascription of beliefs and assertions expressed in the text toward their sources). During the annotation of implicit relations, the annotators inserted a connective expression conveying most closely the meaning of the connection. Where no appropriate implicit connective could be provided, the annotators could use three distinct labels (Prasad et al., 2008, p. 2963): *AltLex* for alternative lexicalizations of discourse connectives like *that is why*; *EntRel* (entity-based relation) for cases where only an entity based coherence relation could be perceived between the segments and *NoRel* (no relation) for cases where none of the relations listed above could be perceived. A closer description of the use of these annotation labels in the PDTB is given in Chapter 12.

The Prague approach to discourse relations is also an annotation-oriented conception. As such, it is inspired by the Penn Discourse Treebank in particular in the following three points:

- definition of a discourse relation,
- the strategy of identification of discourse connectives as pointers to discourse relations in a text, and
- some features of the semantic classification of discourse relations.

2.3 Discourse Connectives

Discourse connectives play an important role in identifying and describing discourse relations since they are the most apparent pointers to discourse structuring on the surface, both for humans and machines. In the Prague approach, the category of discourse connectives is delimited functionally: It contains language expressions whose function is to connect pieces of discourse into a meaningful whole.[9] Discourse connectives (henceforth also DCs) include devices operating both between sentences and within them, cf. *then* and *when* in Example 1 above for the two respective cases. Following the PDTB, we define a discourse connective as a predicate of a binary relation that takes two discourse units as its arguments. A discourse connective combines these units into larger ones, signaling a semantic relation between them. In the Prague annotation scenario, primary and secondary connectives are distinguished (Rysová and Rysová, 2014). The core part of the category, the *primary connectives*, are frequent, mostly one-word expressions that are in principle morphologically inflexible and that usually do not act as grammatical constituents of a sentence. Like sentence modality markers, they are "above" or "outside" the proposition. For details on primary connectives, see Chapter 10. On the other hand, *secondary connectives* are mainly multiword, non-grammaticalized phrases. They are a very heterogeneous class of expressions functioning as sentence elements (like *because of this*), sentence modifiers (*simply speaking*) or even forming separate sentences (*The condition is clear.*). For a more detailed characteristics of this group, see Chapter 11.

Whether a given expression is a discourse connective or not always depends on the particular context. For some expressions, the function of a discourse connective is typical (e.g. *protože* [*because*], *však* [*however*]), other become discourse connectives only in certain contexts (*jinak* [*otherwise*], *podobně* [*similarly*], *naproti tomu* [*on the contrary*, lit. *opposite to_this*], etc.).

Primary connectives are represented by different part-of-speech classes in our approach. According to the part-of-speech (PoS) tagging scenario used for the Prague Dependency Treebank, discourse connectives are represented by the following PoS categories.

– coordinating conjunctions: *a* [*and*], *ale* [*but*], *však* [*but*], *nebo* [*or*], *proto* [*therefore*] ...
– subordinating conjunctions: *ačkoliv* [*although*], *když* [*when*], *místo, aby* [*instead*] ...
– particle expressions: *ovšem* [*however*], *zkrátka* [*in short*], *dokonce* [*even*], *také* [*also*], *například* [*for example*] ...
– some adverbs: *potom* [*then*], *následně* [*afterwards*], *stejně* [*equally/alike*], *současně* [*at the same time*], *tak* [*so*], *totiž* [roughly *because*, *since*, *actually*, *in fact*] ...
– elements formed by letters or numbers expressing enumeration: *a), b), 1., 2.* ...
– two punctuation marks: colon and dash.

As this list indicates, also some punctuation marks can have the function of discourse connectives in certain context, cf. the colon in Example 2.

(2) *Hospodaření Telecomu za rok 1993 není špatné**:** Výnosy činily přes 16 miliard korun, náklady byly přes 11 miliard.* (PDT)

[9] 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.

> *The financial performance of Telecom for the year 1993 is not bad: Revenue totaled over 16 billion Czech crowns, expenses were over 11 billion.*

A detailed PoS and further morphosyntactic characteristics of discourse connectives annotated in the PDT is the topic of Chapter 10.

2.4 Discourse Arguments

Before discussing the units of discourse, we will clarify our use of some syntactic terms. A *clause* is a simple syntactic unit with one predication whereas a *sentence* is understood as a hyperonymous term designating a clause, a compound sentence and also an utterance (a corpus instance).

As already indicated in the analysis of Example 1 above, the two discourse units building a discourse relation are referred to in the present monograph as *discourse arguments*. The Prague annotation scenario also shares the basic notion of a discourse argument with the PDTB, namely the concept of *abstract objects* by Asher (1993). Semantically, abstract objects can be seen as various propositions, i.e. assertions about some set of entities (events, states, situations, facts, beliefs, questions, etc.). Syntactically, in the theoretical view, several constructions can be interpreted as abstract objects. It is mostly individual clauses (the most typical discourse argument is a single clause with a finite verb), connection of clauses, a (compound) sentence, sequences of more sentences, but also deictic expressions referring to previous explicit propositions, nominalizations of clauses, participial and infinitive constructions etc. In annotation practice, the projects aiming to mark large datasets had to restrict the annotation of abstract objects to a manageable subset. Mostly, discourse units (abstract objects) represented by clauses with finite verbs and partially some infinitive and participial constructions are annotated. This is also the case of the Prague discourse annotation. In addition, some elliptical constructions (with elided governing verb) were annotated in the PDT (cf. Poláková et al., 2012a).

In accordance with the PDTB annotation approach, the extent of a discourse argument in the PDT respects the *minimality principle* (Prasad et al., 2007, p. 14), which states that a discourse argument includes only the 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 discourse annotation in Prague, the minimality principle applies mostly to the number of sentences included in a single argument. Dependent clauses (and also the relative ones) within one sentence were mostly considered as a part of the argument. Removing a relative clause from an argument had to be justified.

2.4.1 Notation 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 argument is marked Argument 1 (Arg1). In the Prague annotation, on the other hand, the arguments have been defined semantically. So, for instance, in the relation *reason–result*, the text span expressing the reason is always marked Arg2, and the text span expressing the result is always marked Arg1, regardless of which one contains the connective or in which order they appear in the text. An important annotation rule is that the discourse link (represented by an arrow in the annotation, cf. Figure 2.1) always leads from Arg2 to Arg1. Because of the semantic labeling of the arguments (represented by the oriented discourse link) in the PDT, the Prague repertoire of discourse semantic types could be reduced compared to the PDTB without loss of information, cf. the subsection on semantic types below.

Throughout this book, discourse arguments in the examples taken from the PDT annotation are highlighted with angle brackets and abbreviations: <Arg1:> and <Arg2:>. A discourse connective, if present, is printed in bold. The type of the discourse relation is signaled by a subscript either with the connective (cf. Example 3) or between the arguments (cf. Example 8).[10]

(3) <Arg1: *Poslední statistické sčítání dopravy proběhlo v roce 1990.*> <Arg2: *Za poslední tři roky se* ***však***$_{\textbf{opposition}}$ *na českých silnicích zvýšil provoz.*> (PDT)

<Arg1: *The latest statistical traffic census took place in 1990.*> <Arg2: *Over the past three years,* ***however***$_{\textbf{opposition}}$, *traffic on Czech roads has increased.*>

Figure 2.1 presents the way annotation of discourse relations was carried out in the PDT for Example 3.[11] The discourse relation of *opposition* is represented by an orange arrow between the root nodes *to take place* and *to increase* of the two arguments. The arrow always points from Arg2 to Arg1. In this way, it can capture the different nature of the arguments for certain types of relations.

2.5 Semantic Types of Discourse Relations

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 monograph, we use the term *senses* only when referring to the PDTB annotation scheme and categories.

[10] In our approach, a connective is not a part of any of the arguments. However, for easy reading of the examples in this book, a connective that is syntactically incorporated into one of the arguments is kept within the argument brackets. Wherever possible, the connective is placed outside the argument brackets.

[11] The English translations of the Czech lemmata in the tectogrammatical trees are not part of the treebank data. The translations have been added to the trees in the figures in this book for easier comprehensibility.

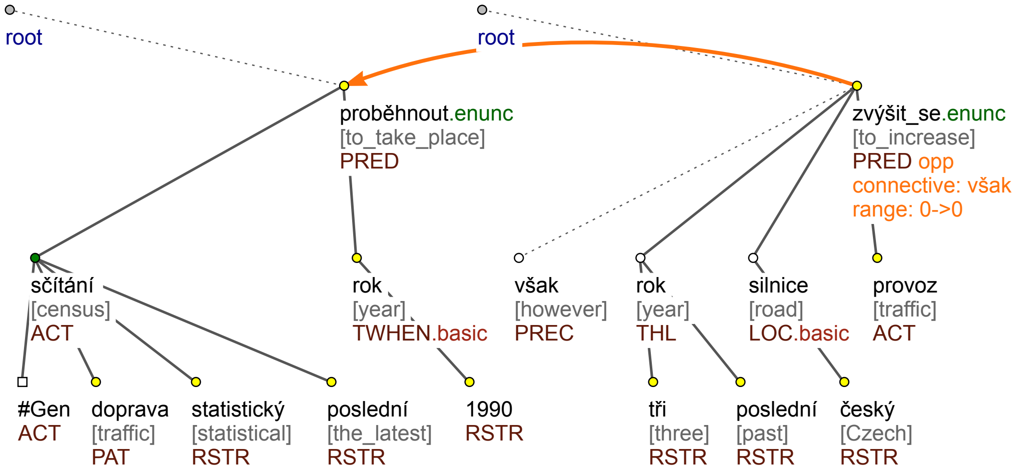


Figure 2.1: Discourse annotation of Example 3

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 in the Prague Dependency Treebank, TEMPORAL, CONTINGENCY, CONTRAST and EXPANSION are identical to those in the PDTB[12] but the hierarchy itself has only two levels, with a total of 22 relations. The third level of the Penn hierarchy is captured by the direction of the discourse arrow (as stated earlier). Within these four classes, the types of relations partly differ from the PDTB types and go closer to Prague tectogrammatical functors. The discourse-semantic categories for the annotation in the PDiT 1.0 and the PDT 3.0[13] are presented in Table 2.1.[14]

We believe that language-specific features can slightly influence a fine-grained semantic classification (cf. Mladová et al., 2009). The semantic classification of discourse relations in the Prague annotation, compared to the PDTB 2.0 label set, was extended by five categories. In the CONTINGENCY class, it is the categories of *purpose* (Example 4), based on the traditional syntactic category – modification of purpose, and *explication* (Example 5), in which the second argument in the linear order typically

[12] With one terminological exception: The COMPARISON class is referred to as CONTRAST class in the Prague scheme.

[13] The Prague Discourse Treebank 1.0 (PDiT 1.0), a predecessor of the Prague Dependency Treebank 3.0, contains the first publicly released discourse annotation, cf. Section 2.8. There were no adjustments of the semantic classification from the PDiT 1.0 towards the PDT 3.0.

[14] In both published versions of the annotated data (PDiT 1.0 and PDT 3.0), older abbreviations for *pragmatic reason–result*, *pragmatic condition* and *pragmatic contrast* were used (*f_reason*, *f_cond* and *f_opp*, respectively).

Name of the relation	Label
TEMPORAL	
synchrony	*synchr*
asynchrony (precedence–succession)	*preced*
CONTINGENCY	
reason–result	*reason*
pragmatic reason–result	*p_reason*
explication	*explicat*
condition	*cond*
pragmatic condition	*p_cond*
purpose	*purp*
CONTRAST	
confrontation	*confr*
opposition	*opp*
restrictive opposition	*restr*
pragmatic contrast	*p_opp*
concession	*conc*
correction	*corr*
gradation	*grad*
EXPANSION	
conjunction	*conj*
conjunctive alternative	*conjalt*
disjunctive alternative	*disjalt*
instantiation	*exempl*
specification	*spec*
equivalence	*equiv*
generalization	*gener*

Table 2.1: Semantic types of discourse relations in the PDiT 1.0 and the PDT 3.0

gives a non-causal clarification, or explanation of the first one. In the CONTRAST class, three new discourse-semantic types were introduced, two of them in order to sub-classify a more general adversative meaning (for details cf. Chapter 9): *Restrictive opposition* (Example 6), which also includes the meaning of exception, *gradation* (Example 7) and *correction* (Example 8).[15]

[15] For the relation of *correction*, a negative expression in the preceding context is obligatory. This relation is typically expressed by the Czech connective *nýbrž* [*but; not x – but y*] which corresponds to the German expression *sondern*.

(4) <Arg1: *Chystáme snížení množství oprav na poštovních budovách,>* <Arg2: ***abychom***$_{\text{purpose}}$ *ušetřili.>* (PDT)

<Arg1: *We plan to reduce the amount of repairs to the postal buildings>* ***in order to***$_{\text{purpose}}$ <Arg2: *save money* (lit. ***in_order_that_we*** *save*).>

(5) <Arg1: *Nejen doping odvádí pozornost od sportovních výkonů.>* <Arg2: *Kanadská policie* ***totiž***$_{\text{explication}}$ *pátrá po sedmi reprezentantech, kteří v průběhu her opustili atletickou vesnici a zdržují se na neznámém místě.>* (PDT)

<Arg1: *Not only doping diverts attention from the athletic achievements.>*
As a matter of fact$_{\text{explication}}$, <Arg2: *the Canadian police are looking for the seven athletes who left the Olympic village during the games and are staying at an undisclosed location.>*

(6) <Arg1: *Každá krajina má svou krásu.>* ***Jenom***$_{\text{restr. opposition}}$ <Arg2: *ji musíte umět vidět.>* (PDT)

<Arg1: *Every landscape has its beauty.>* ***Only***$_{\text{restr. opposition}}$ <Arg2: *you must be able to see it.>*

(7) <Arg1: *Sabotage bodovala* ***nejen*** *v rodné Americe,>* ***ale***$_{\text{gradation}}$ <Arg2: *pronikla* ***i*** *do žebříčků evropských.>* (PDT)

<Arg1: *Sabotage topped the charts* ***not only*** *in America,>* ***but***$_{\text{gradation}}$ <Arg2: *it* ***also*** *made it onto the European charts.>*

(8) <Arg1: *Stát* ***ne****ní soukromým majetkem ústavních orgánů.>*$_{\text{correction}}$ <Arg2: *Je to veřejnoprávní instituce.>* (PDT)

<Arg1: *The state is* ***not*** *private property of the constitutional bodies.>*$_{\text{correction}}$ <Arg2: *It is a public institution.>*

One of the most discussed properties of discourse relations is their "*semantic*" or "*pragmatic*" nature, in other words, the question of what is actually related – propositions, inferences, illocutions, etc. This distinction is a little confusing, as the relations are always semantic but they either hold between text contents or between the inferred materials.[16]

In the PDTB, four pragmatic senses are distinguished and annotated: pragmatic cause, condition, contrast and concession. In the Prague scenario, three pragmatic meanings were annotated. *Pragmatic concession* and *pragmatic contrast* were merged into a single group for the lack of reliable distinctive features. Example 9 demonstrates the relation *pragmatic reason–result*. In this example, there is no causal relation between the fact that the qualification for the European Championship in football has already

[16] The issue of the distinction between the notions *semantic* and *pragmatic* is addressed in more detail in Poláková (2015).

started and providing the overview of the teams' football history. Rather, the authors of the article justify their choice of topic by citing the current affairs of European football.

(9) <Arg2: *Zatímco většina fotbalových reprezentací vstupuje do kvalifikace pro ME 1996 nyní v září, boj o účast v Anglii vypukl již dříve.*> (...) <Arg1: *Před opravdovým rozjezdem kvalifikace* ***proto***$_{\textbf{pragm. reason–result}}$ *přinášíme přehled, jak často spolu celky v jednotlivých skupinách už v soutěžích ME a MS v minulosti hrály.*> (PDT)

<Arg2: *While most national football teams enter the qualification for the 1996 European Championship now, in September, the fight for a place at the competition in England started earlier.*> (...) <Arg1: *Before the real start of the qualification, we* ***therefore***$_{\textbf{pragm. reason–result}}$ *provide an overview of how often the teams in each group had played each other at European and World Championships in the past.*>

2.6 Annotation Process

The present section provides a brief description of the process of the build-up of the Prague Discourse Treebank 1.0 (PDiT 1.0). Throughout this section, 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 in the Prague Treebank family. Where needed, we describe the adjustments and changes in the newer data release, the PDT 3.0.

2.6.1 Theoretical starting points

One theoretical issue related to the nature of discourse relations is of particular interest for us – the (partial) correspondence of discourse structure and semantics to the structure and semantics of a sentence. We mention it here for the sake of completeness; Chapter 9 addresses this topic in more detail. The fact that the discourse project in Prague is based on the previous annotation of underlying syntax reflects the basic assumption that, from the cognitive viewpoint, the semantics within a sentence is the same as the semantics of discourse relations. 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 (Jínová, Poláková and Mírovský, 2014).

When we analyze 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 ascertain that discourse relations, or at least some of them, are syntactically motivated and syntax-bound: When we cross the sentence boundary, we find the same semantic patterns. From the opposite point of view, when we start analyzing discourse composition, we will sooner or later arrive at discourse-relevant intra-sentential phenomena. Thus, there is no doubt that sentence syntax and semantics are of great relevance to discourse analysis.

This fact constitutes a basic starting point for the project: A syntactico-semantic analysis of a sentence contains (retrievable) information about relations in discourse. Annotation of discourse relations in Czech was therefore quite a natural step: We had at our disposal a large, multilayer-annotated resource for Czech (the PDT 2.5), the tectogrammatical layer of which already offered some information potentially relevant for discourse annotation in the sense of the PDTB. The two main decisions for the representation of discourse relations in the Prague scenario, namely the inspiration from the PDTB annotation scenario and the decision to annotate discourse relations directly on top of the tectogrammatical trees are discussed in the following two sections.

2.6.2 Inspiration from the PDTB approach

The approach of the PDTB group was reflected in the build-up of the PDiT 1.0 in two main ways: The first is the basic concept of connective identification, finding the two arguments of the connective and assigning a semantic label to the relation signaled by the connective. The second point that was borrowed from the PDTB was the shape of the hierarchy of sense tags for discourse. In the PDiT 1.0 and the PDT 3.0, the annotations of discourse relations are limited to the relations expressed by explicit discourse connectives (present on the surface); other tags (for implicit relations, AltLex, EntRel and NoRel in the original PDTB sense)[17] between adjacent sentences were not assigned. Alternative lexicalizations (AltLex) are treated in a more complex way, as part of an extensive analysis of secondary connectives. Their annotation was carried out in a later phase of the project. For details see Chapter 11. 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 and PDT 3.0 as a part of another subproject (cf. Chapters 3 and 4). Another phenomenon not annotated in Prague treebanks so far, compared to the PDTB, is attribution. We believe that this information can be partially obtained from syntactic features of the syntactic layers of the PDT, e.g. attributes for direct speech, parentheses, verbal valency etc. (Poláková et al., 2013).

2.6.3 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 manually annotated information that can be relevant for the representation of discourse structure. This is quite a unique approach among the similarly aimed

[17] See Chapter 12, Sections 12.1.2 and 12.1.3.

projects and it brings many (both linguistic and technical) advantages. The main benefits are the easy retrieval of intra-sentential discourse relations and their connectives, resolved elliptical structures, marking of parentheses, reporting clauses, appositions, coordinations of mere noun phrases etc. The possibility for the annotator to search for and visualize more linguistic phenomena at once was also of great advantage. A detailed look on the mutual relationship of syntactic and discourse analyses in the PDT is given in Chapter 9.

2.6.4 Two-phase annotation

A rather practical decision resulted from the intention to annotate discourse directly on top of syntactic trees – to proceed in two annotation phases. In the first phase, the treebank was thoroughly manually annotated with a focus on inter-sentential discourse relations (relations between sentences) signaled by explicit discourse connectives. Intra-sentential relations were only marked manually in the cases where the tectogrammatical representation did not convey a certain type of discourse semantics, according to the annotation guidelines. 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. Both types of annotation underwent consistent checking procedures (cf. Chapter 7).

The 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 the PDTB group used, where an annotator went through all the occurrences of one connective type in the whole treebank. This way, the set of possible discourse connectives 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 discourse connective, according to the criteria for discourse connectives set in advance in the annotation guidelines. Thus, the question whether a certain expression in a certain context actually fulfills the criteria of a discourse connective could arise. Also, the annotators were free to mark more expressions as connectives of a single relation, which allowed them to capture many modified connectives (*právě protože* [*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 distinguishing whether a co-occurrence of more connective expressions means that they signal a single discourse relation or more. This approach may be less consistent as for the delimitation of the category of discourse connectives, but it does point to some interesting linguistic material on the periphery of this category and enables its further research.

Only after having searched for discourse connectives in the hard copies of the corpus texts, the annotators worked with the tree structures in the TrEd annotation tool (cf. Section 8.3.1). Having identified the connective, its two arguments (i.e. their extent) were set (creation of the discourse arrow), and one of the labels for semantic types was assigned to each of the relations.

Another difference in the process of annotation in Prague, in contrast to the PDTB, was the assignment of semantic labels (sense tags) to the relations. The PDTB annotators were not forced to make the finest distinction (on the subtype level). In the Prague semantic label assignment, on the other hand, the annotators had to choose one of the 22 types.

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, Mírovský and Poláková, 2012b and also Chapter 9 of this monograph).

The 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 process of transfering syntactico-semantic labels (functors) to discourse semantic types is described in greater detail in Chapter 9.

Unlike tectogrammatical relations, discourse-semantic relations in our approach do not reflect syntactic subordination and coordination. These two basic formal principles of grammatical arrangement of a sentence are, in particular in the European approaches to syntax, strongly connected to certain semantics. For instance, the meaning of condition is typically connected to the subordinate 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 claim that the semantic types introduced for discourse mostly have both possibilities of expression. In our conception, discourse relations can be expressed via syntactic subordination or coordination within a single sentence, and further between individual sentences or larger text units. Example 10 demonstrates a conditional meaning expressed by coordinating means. Thus, the syntactic distinction of subordinate and coordinate structures does not play a role in the design of our semantic classification for discourse.

(10) <Arg2: *Posluchač musí přistoupit na pozici, že vše je dovoleno.*> ***Potom***$_{\textbf{condition}}$ <Arg1: *se pobaví a také pochopí, že drama znázorňuje ztrátu reálné komunikace.*> (PDT)

<Arg2: *The listener has to accept the fact that everything is permitted.*> ***Then***$_{\textbf{condition}}$ <Arg1: *he can enjoy himself and also understand that the drama symbolizes the loss of a real-life communication.*>

2.7 Other Annotated Phenomena

Apart from discourse relations, several other discourse-related phenomena have been annotated in the PDT 3.0. It is rather smaller and less frequent phenomena that nevertheless play some distinctive role in discourse structuring. These phenomena include list structures (cf. Section 2.7.1) or specific discourse-structuring signals like headings and photo captions (2.7.2). Separately, within a later project, manual annotations of genres of the treebank documents were added (2.7.3).

2.7.1 List structures

List structures are enumerative constructions, annotated in the PDiT 1.0 and in the PDT 3.0 as independent compositional structures. A list structure in the Prague approach does not have a semantic label in the semantic types hierarchy, as it is the case in the PDTB annotation. It is annotated as a separate phenomenon for two reasons: First, in this type of structure, every item of the list is related both to the preceding item and to the (facultative) introductory statement for the whole list, if present. The nature of a list structure is therefore 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 filling. In our viewpoint, there is only a *specification* relation between the hypertheme (introductory statement) and the set of list items. 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 linking 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 11. The first sentence is the hypertheme; the connectives are points 1 and 2.

(11) *K tomu, aby zaměstnavatel pracovníkovi za škodu opravdu odpovídal, musí být splněny tyto podmínky:* **1.** *Zaměstnanci musí vzniknout škoda, tj. musí dojít k určitému 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.* (PDT)

> *So that the employer is truly responsible for the damages caused to an employee, the following conditions must be satisfied:* **1.** *The employee must incur damages, 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 physical or legal entity acting on his behalf, must violate their legal obligations.*

2.7.2 Discourse special: headings, captions, metatexts

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 the corpus compilation. 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 in our data. The other two possible values of the *discourse_special* attribute are also incorporated in the PDT 3.0 in the newly added full-scale annotation of genres of the corpus texts.

2.7.3 Genre annotation

Inspired by studies on genre distinction and classification (e.g. Webber, 2009; Taboada, Brooke and Stede, 2009), we carried out manual annotations of the PDT texts for their genres. The 3,156 documents of the PDT previously annotated for tectogrammatics and discourse phenomena were manually assigned a simple label according to their genre characteristics within the journalistic domain (Poláková, Jínová and Mírovský, 2014). *Caption* and *metatext* are among the 20 distinguished genre categories, although not among the most frequent ones. The complete list of the assigned genre categories is given in Table 2.2.

2.8 Summary

This chapter describes the Prague approach to the analysis of discourse relations as one aspect of discourse coherence. It is a lexically based, shallow discourse model focused on the identification of discourse connectives as anchors of discourse relations. Similar to the Penn Discourse Treebank – a leading project in this research field – our approach is oriented on a large-scale corpus annotation. The annotation of discourse relations, arguments and connectives in Czech was carried out on almost 50,000 Czech sentences and first published in 2012 under the name Prague Discourse Treebank 1.0 (Poláková et al., 2012b). Its enhanced version (including genre annotation) is a part of the most recently released Prague Dependency Treebank 3.0 (Bejček et al., 2013).

Monologic genres	Dialogic genres	Other
critical review	*topical interview*	*collection*
invitation	*personality-focused interview*	*caption*
letters from readers		*metatext*
advice column		*other*
cultural program		
film/TV program		
sports news		
comment		
news report		
reflective essay		
overview		
description		
weather forecast		
readers' survey		

Table 2.2: Genre categories in the PDT 3.0

3

Coreference

As discussed in the previous chapters, textual coherence is achieved through various types of conceptional and cognitive relations created in the text. Discourse relations addressed so far (see Chapter 2) hold between predicative elements, such as clauses, sentences and larger textual segments.[18] We will now discuss relations between non-predicative items, primarily between nominal groups. Take a look at Example 12:

(12) *John asked his mother to advise him on how he should behave with Mary. She ignored her son's wish.*

As we can see, there is no explicitly expressed discourse relation between the two sentences. However, the sequence remains coherent due to the implicit relation of confrontation and the coreferential relations between entities. In the given example, the following elements are coreferential (i.e. they relate to the same discourse entity):

- *John – his – him – he – her son*,
- *his mother – she – her*.

The relation between *to advise him how he should behave with Mary* and *wish* is partly different, as it holds between infinitive clause and a deverbative noun. However, this is still the relation of identity of discourse segments, and in this respect it is closer to coreference than to discourse relations as described in Chapter 2. Such relations are often referred to as *discourse deixis* and will be also addressed within the notion of coreference in this chapter.

Inspired by Paducheva (1985) and Langacker (2008), we understand coreference as the relation holding within the world of discourse. In other words, we establish coreferential relations between entities realized in the utterance (not between word meanings as it was common in classical logical semantics), thus the existence of respective objects in the real world is not essential.

3.1 Basic Terms

The first notion implied by the term *coreference* is identity of referents signified by language expressions in discourse. Thus, the phenomenon of coreference is primarily

[18] It is also possible to speak about discourse relations between nominal groups but in this case they should have predicative function (nominalizations, gerunds, infinitives etc.).

based on *reference* and *identity of referents*. Further, coreference is close to the notion of *anaphora*. We will discuss these terms in more detail.

Reference. In linguistics, *reference* has two different interpretations. On the one hand, it is the ability of language expression to refer to discourse entities, which may be further linked to extralinguistic objects. From this point of view, we distinguish between *specific* and *generic reference* (there is rich linguistic literature on this topic, e.g. Carlson and Pelletier, 1995; Hlavsa, 1975; Palek, 1968; Paducheva, 1985). Generic reference takes place by any member representative of a class of entities (see e.g. *the dog* and *the jackal* in Example 13) and specific reference points on a particular specimen of the class (e.g. the nominal group *my dog* in Example 14).

(13) ***The dog*** *has a common ancestor with* ***the jackal****.*

(14) ***My dog*** *is very old.*

On the other hand, the term *reference* is also used to name textual links to preceding or following context, eventually also referring out of the text, to extralinguistic circumstances. From this point of view, exophoric and endophoric reference are distinguished. *Exophoric reference* or *exophora* is referring to a situation or entities outside of the text. So, in Example 15, the nominal group *this week* refers to the actual time when the utterance is made, it has deictic meaning. *Endophoric reference* is referring to elements within the text, as shown in Example 16. Here, the nominal group *this week* refers to the time distance *between September 25th and 30th* mentioned in the previous sentence.[19]

(15) ***This week****, the workshop on discourse annotation is being held in Prague.*

(16) *Peter planned to have his exams* ***between September 25th and 30th****.* ***This week*** *was the last possibility to finish the academic year.*

Identity of referents. Understanding coreference as the *identity of referents*, with the assumption that the notion of identity is considered to be a dichotomy of identity and non-identity is quite problematic. First of all, identity itself is not a trivial notion. Take a look at Example 17. The identity of *the Gipsies* and *this nation* is very likely but it does not stay to reason. The first sentence refers to certain Gipsies during the second world war. The second sentence speaks in general about the whole nation. However, we are inclined to consider these groups to be definitely coreferential.

(17) *Nic z toho se však nevyrovná míře neštěstí, které* ***Romy*** *postihlo v letech druhé světové války. Spolu se Židy byli označeni za méněcennou rasu a stali se objektem patologických fašistických opatření, jejichž cílem byla úplná genocida* ***tohoto národa****.* (PDT)

[19] One can also imagine an exophoric interpretation of *this week*. However, in this case, the text will be incoherent.

> *Nothing of this, however, compares to the misfortune that befell* ***the Gipsies*** *during the Second World War. Together with the Jews, they were called an inferior race and became the object of pathological fascist measures, their purpose being the complete genocide of* ***this nation****.*

Similarly, the identity of entities (persons, cities, relations, etc.) in different periods of time requires a big portion of human imagination. This uncertainty gave birth to the introduction of the concept of *near-identity* (Recasens, Hovy and Martí, 2010). The authors present the concept of identity as a scale ranging from obvious identity to obvious non-identity with a large range of cases inbetween. They further classify near-identity relations into different types, such as spatio-temporal function, meronymy, different kinds of metonymy, etc. This approach is argued in Ogrodniczuk et al. (2015). The authors believe that most of the near-identity types can be explained by various phenomena on the levels of grammar, semantics and concepts. However, coreference is a property of the discourse world and it is realized on the pragmatics level only. To produce and understand coherent texts, the information about grammar, semantics and real world knowledge can be used but it is not indispensable, because texts function on the discourse level and the coreference relations are interpreted within it. Thus, introducing an additional term of near-identity "... does not explain anything, and rather disturbs the structure of annotated texts, as it mixes up separate levels of language – system and speech" (Ogrodniczuk et al., 2015, p. 22).

Coreference and anaphora. The phenomena of coreference and anaphora are very complex, closely interrelated and often differently understood depending on the scholar and scientific conception. Therefore, these concepts can be very easily confused.

Both anaphora and coreference are basic means of achieving text cohesion (see e.g. Halliday and Hasan, 1976; Palek, 1968; Prasad et al., 2008; Hrbáček, 1994). *Anaphora* is an inter-textual system; it is the relation of an *anaphoric expression* (or *anaphor*) to a textual item that has been mentioned in the previous context (*antecedent*), where the correct interpretation of the anaphoric expression depends on the antecedent. In this respect, anaphora is opposed to *cataphora*, that is a reference to a so-called *postcedent* in the following context in the text. On the other hand, *coreference* is a referential identity of language expressions, which means that two or more elements in the text refer to the same phenomenon in the world of discourse.

Coreference and anaphora often occur simultaneously (see e.g. the relation between *mother* and *she*, or *John* and *him* in the introductory Example 12), but this is not always the case. For example, *a book* and *one* in Example 18 are anaphoric but not coreferential. On the contrary, the instances of *Prague* in Example 19 are coreferential but not anaphoric, because the noun *Prague* in the first sentence is not needed for the correct interpretation of *Prague* in the second sentence.

(18) *Peter has* ***a book****. Mary also has* ***one****.*

(19) *I like* ***Prague****.* ***Prague*** *is one of the most beautiful cities in Europe.*

Our main concern in this chapter is coreference; we are interested in language expressions referring to the same discourse entity within the text. This relation is originally not directional, it only refers to the identity of referents. However, we study coreference in written texts which are linear, so we want to use this structuring in our analysis. Therefore, for the sake of convenience and in order not to be forced to invent new terms, we will use the terms from the field of anaphora, i.e. the notions of *anaphor* or *anaphoric expression* for language items that are coreferential with some expression in the preceding context and the notion of antecedent for coreferred expressions. In the case of cataphoric reference to the following context, the notions of *cataphor* and *postcedent* are used, respectively.

Language expressions referring to the same discourse entities may be observed either as sets (or clusters) of coreferential expressions or as *coreference chains*. In our approach, we consider coreferential expressions to be organized in chains, the first mention of a chain being mostly the antecedent (except for some rare cases of cataphoric reference). Considering the text structure and taking text cohesion into account, it is often hard to decide to which specific antecedent the given anaphoric mention refers. Coreference chains are not always simple and straightforward, they do not have to connect one expression to another. They can also split, when one anaphoric expression summarizes more than one antecedents. See Example 20, where the anaphoric pronoun *their* refers simultaneously to three antecedents: *she*, *her children* and *her war-damaged husband*.

(20) *Although* ***she*** *was kind and playful to* ***her children****, she was dreadful to* ***her war-damaged husband****; she openly brought her lover into* ***their*** *home.*

(Wall Street Journal, PEDT)

3.2 Related Work

The phenomenon of coreference is usually mentioned in the context of text coherence and it is closely related to the theory of reference and anaphoric studies. Furthermore, during the last few decades, coreference is significantly explored in computational linguistics.

Coreference is an effective means of text coherence, making it possible to unify the message consisting of a set of clauses into a coherent whole. This aspect is addressed mainly in theoretical linguistics, in the theory of communication, and especially in studies analyzing cohesion and coherence (see Halliday and Hasan, 1976; Hobbs, 1979; Kehler et al., 2008, etc.). In this field, the mechanisms of pronoun interpretation are studied, the reasons for different interpretations being explained first of all by semantics, inference and psycholinguistics.

The theory of reference originates in logical semantics (Frege, 1892; Russel, 1905; Carnap, 1947, etc.). It examines the relation of language expressions to the entities in the real world. From this point of view, the abilities of language expressions to refer are explored. The sets of referential properties are different in different approaches, and the extent of such sets also depends on the goal of the analysis. The distinction between referential and predicative uses of noun phrases (cf. ***The doctor*** *came* vs. *My brother is* ***a doctor*** resp.) is accepted by almost all studies addressing this topic, while further specifications are rather diverse. There is a number of studies addressing the notion of genericity, generic noun phrases and their ability to refer. According to different researchers, generic nominal groups are considered to be either referring to classes (e.g. in Carlson and Pelletier, 1995 and Mendoza, 2004) or non-referring (rather predicating) classifications (Paducheva, 1985), which are able to have specific and non-specific interpretations (in Mendoza, 2004 and Shmelev, 1996) and to be distinguished from non-specific nominal groups as a separate type (Carlson and Pelletier, 1995; Paducheva, 1985). Some researchers (Paducheva, 1985; Adamec, 1980) single out intermediate reference classes, such as existential reference (*It would be nice, if you marry* ***a foreigner***), distributive reference (***Each*** *of us visited him*), relatively specific reference (*I'd like to see* ***an interesting film***) and so on.

Further, coreference and coreference resolution (automatic search for a proper antecedent of nominal groups in text) is one of the core research topics in computational linguistics. There is a number of anaphorically annotated corpora for different languages. One of the most detailed annotations of coreferential links is carried out in the project MATE and in the related projects GNOME and VENEX, focused on English and Italian (Poesio et al., in prep.; Poesio and Artstein, 2008), while a detailed annotation of coreferential links with a profound analysis of the acquired data is provided for Spanish and Catalan (Recasens and Martí, 2010). Among annotated corpora of Slavic languages, there is a very systematic coreference corpus of Polish (the CORE project; Ogrodniczuk et al., 2015), German-English Contrasts in Cohesion (the GECCo corpus) for German and English (Lapshinova-Koltunski and Kunz, 2014) and the RuCor for Russian (Toldova et al., 2014). Annotated corpora of the kind mentioned above serve as useful data for systems of automatic recognition of coreferential links in texts (Lee et al., 2013), for the determination of degrees of salience (Poesio, 2003), statistical models of language generation (Cheng et al., 2001) and other similar tasks in the domain of automatic NLP and information retrieval.

3.3 Grammatical and Textual Coreference

In Czech linguistics, grammatical and textual coreference have been traditionally distinguished (Hajičová, Panevová and Sgall, 1985; Hajičová, Panevová and Sgall, 1986; Hajičová, Panevová and Sgall, 1987). *Grammatical coreference* is associated with the syntactic structure of sentences, it is grammar-driven by the use of pronouns and, in most cases, it is possible to identify the antecedent on the basis of grammatical rules. In the case of grammatical coreference, both antecedent and anaphor are located in the same sentence (some exceptions can be found in Hajičová, Oliva and Sgall, 1987). *Textual coreference* is not restricted to grammatical means alone. It can be realized by pronominalization, grammatical agreement, repetitions, synonyms, paraphrasing, hyponyms/hyperonyms, etc. Unlike grammatical coreference, textual coreference often occurs between entities in different sentences. The distinction between grammatical and textual coreference is considered to be basic and thus we will consider them separately in Sections 3.3.1 and 3.3.2.

3.3.1 Grammatical coreference

In this chapter, we present a brief survey of cases that can be considered as grammatical coreference. The topic and typology of grammatical coreference in the Czech anaphoric approach have been addressed in more detail in Hajičová, Panevová and Sgall (1985); Hajičová, Panevová and Sgall (1987); Panevová (1991); Hajičová, Oliva and Sgall (1987), etc.

There are two possible ways of expressing grammatical coreference: either anaphor has the form of a pronoun or it is given by the syntactic structure of the sentence, thus not being expressed on the surface level but reconstructed on the tectogrammatical level (see the description of the tectogrammatical layer in Chapter 6, Section 6.1).

The following types of grammatical coreference can be distinguished:

1. Coreference with reflexive pronouns and the Czech possessive reflexive *svůj*. In this case, the anaphoric pronoun mostly refers to the closest subject, cf. Example 21, where the reflexive pronoun *sobě* corefers with the subject *matka* [*mother*], which corresponds to the *Actor*[20] argument of the verb *přát* [*to let_have*].

(21) ***Sobě** nedopřeje **matka** nikdy nic.* (Mikulová et al., 2005)
lit. ***To_herself** not_let_have **mother** never nothing.*
Mother never treats herself to anything pleasant.

This is also the case with the possessive reflexive *svůj* [*his*, *her*, *its*] in Czech,[21] but with some exceptions. In the cases when the reflexive *svůj* is used in clauses with the third-person predicate, it can refer to any argument, including those that are not in the subject position, cf. Example 22, where the possessive reflexive *svůj* corefers with the indirect object *jim* [*them*].

(22) *Jejich kajakářské disciplíny oplývají desítkami vynikajících soupeřů a je také pravděpodobné, že při **svém** profesionálním přístupu k závodění **jim**$_{\text{Dative}}$ chybí trochu víc uvolněnosti.* (Mikulová et al., 2005)
Their kayak disciplines have dozens of brilliant rivals, and it is also possible that (lit. *by **self's** professional attitude, **them**$_{\text{Dative}}$ lack the ability to relax*) *with **their** professional attitude, **they** might lack the ability to relax.*

[20] For the meaning of tectogrammatical functors see Section 6.1.

[21] The category of possessive reflexive is missing in English, the Czech possessive reflexive *svůj* is thus translated into English with possessive pronouns *his, her, its, their*, etc.

Example 23 shows grammatical coreference of reflexive *sám* [*himself*]:

(23) *Jak říká **sám pan Bronner**, ve vzduchu byl cítit zápach syrového masa.* (PCEDT)
*As **Mr. Bronner himself** says, the smell of raw meat was in the air.*

2. Coreference with relative elements. Relative pronouns and pronominal adverbs introducing relative clauses are linked to their antecedent in the governing clause, as in Example 24. Here, the relative pronouns *níž* [*which*] and *která* [*which*] corefer with the noun *síť* [*net*] modified by the dependent relative clause.

(24) *Za informační dálnici se považuje světová telekomunikační síť, po **níž** lze přenášet zvuk, data i obraz a **která** tak otevírá přístup k množství informatických služeb.* (Mikulová et al., 2005)

*A net **which** makes it possible to transfer sound, data and picture and **which** opens access to many informational services can be considered to be an information highway.*

3. Control. The relation of *control* is a type of grammatical coreference that arises with certain verbs, called control verbs, such as *begin*, *let*, *want*, etc. The control relation arises, for example, with the elided subject of the infinitive *sleep* and the subject *Peter* in Example 25.

(25) *Peter wants to sleep.* (Mikulová et al., 2005)

This is such a coreferential relation between controller and controllee, that (i) the controller is a member of the valency frame of the governing verb; e.g. in Example 25, *Peter* is a member of the valency frame of the verb *to want*, (ii) the controllee (in our case the elided subject of the infinitive *to sleep*) is a member of the valency frame of the infinitive (*to sleep*) dependent on the control verb and (iii) the infinitive is a valency modification of the control verb; e.g. in Example 25, *to sleep* is a valency modification of the verb *to want*.

The control relation depends on the lexical semantics of the control verb. It appears to be possible to make a list of control verbs, or at least to make a list of verbal meanings that will tend to be expressed by control verbs. These are, first of all, modal verbs (***can** read*), phrasal verbs (***begin** to read*), intention verbs (***plan** to read*) and so on.

The control relation is related to so-called *quasi-control* relation that is a specific grammatical coreference relation that can be found with multiword predicates the dependent part of which is a noun with valency requirements (such as *duty*, *requirement*, *protection*, etc.). The fact that some combinations of noun and verb make common lexical entities causes the referential identity of some of their arguments. In the surface structure, the identical modifications are usually expressed only once; cf. Example 26, where the Addressee of the verb *poskytnout* [*to provide*] as well as the Patient[22] of the

[22] For the meaning of tectogrammatical functors see Section 6.1.

noun *ochrana* [*protection*] has the same reference (*Jan*). This shared modification can only be present once at the surface level (it is impossible to say: **Dan poskytl Janovi ochranu Jana* [lit. *Dan offered Jan protection of Jan*]).

(26) *Poskytl Janovi ochranu.* (Mikulová et al., 2005)
He[23] *offered Jan protection.*

4. Coreference with verbal modifications that have dual dependency. In this case, grammatical coreference concerns non-expressed arguments of verbal modifications with the so-called *dual dependency* (e.g. passive participles, gerunds, infinitives). This is, for example, the case of coreference of unexpressed Actor of the infinitive *běhat* [*to run*] with the Patient *Hanka* of the governing verb *zastihl* [*saw*] in Example 27.

(27) *Honza zastihl Hanku běhat kolem rybníka.* (Mikulová et al., 2005)
Honza saw Hanka run around the lake.

3.3.2 Textual coreference

Textual coreference consists of the cases of referential identity which are not covered by grammatical coreference, i.e. where antecedent cannot be easily resolved using the grammatical rules of a given language.

We speak about *pronominal textual coreference* if anaphoric expression is either elided on the surface level or it is expressed by personal, possessive or demonstrative pronoun. Pronominal textual coreference is primarily anaphoric (rarely cataphoric), in any case there is an endophoric (intra-textual) reference to the preceding (or the following) context. However, endophoric reference is not required in the case of *nominal (extended) textual coreference* that takes place if anaphor is expressed by other means than pronouns or ellipsis.

Let us now discuss which types of language expressions may take part in textual coreference relations. It stands to reason that basic and the most frequent anaphors are nominal groups (nouns, pronouns, nominal demonstratives and nominal groups with the noun head). However, it is often a good idea to consider referential properties of other expressions. In what follows, we will describe different types of language expressions that may take part in coreferential relations according to our approach.

1. Anaphoric zeros. The notion of anaphoric zero stems from the theory of Functional Generative Description (Sgall, 1967b; Sgall, Hajičová and Panevová, 1986). Anaphoric zero is always a textual ellipsis of a dependent element, i.e. the omitted element is a dependent modification (an argument) of its governing expression and it can

[23] The subject is elided in Czech (see description of Czech pro-drop nature in Chapter 1).

be identified from the previous (or, less frequently, the following) context.[24] See Example 28:

(28) ***Umělec*** *má svůj denní řád. ∅ Tráví den kreslením portrétů kolemjdoucích či se o to alespoň snaží.* (PCEDT)

The Artist *has his routine.* ***He***[25] *spends his days sketching passers-by, or at least trying to.*

2. Personal and possessive pronouns. Resolution of personal and possessive pronouns relies on the context and cannot be derived from grammatical rules of Czech. Let us return to the first sentence of the introductory Example.

(29) *John asked his mother to advise him on how he should behave with Mary.*

The most probable interpretation of personal and possessive pronouns used in the example is based on our expectation of text coherence and the presence of explicit antecedent in the immediately preceding context, so coreference chain *John – his – him – he* is expected. However, for example, *his mother* is not obligatorily John's mother, one can also imagine that she can be a mother of some other person mentioned earlier. Similarly, other antecedents for *he* and *him* are possible.

3. Nominal demonstratives. Demonstrative pronouns being used as nouns (i.e. having a denotative function and referring to entities themselves, not as an attribute in the noun phrase) enter into textual coreference relations in the same way as personal and possessive pronouns (Example 30).

(30) *Ta přijala* ***strategii Bílého domu*** *v domnění, že je* ***to*** *nejjistější cesta k vítězství.* (PCEDT)

She endorsed ***the White House strategy****, believing* ***it*** *to be the surest way to victory.*

In Czech, if one needs to refer to a sentence or a longer utterance, the demonstrative pronouns are often used. Similarly as in Example 30, we consider these cases as textually coreferential (Example 31).

(31) *Jako herec* ***není*** *Charles Lane* ***dědicem ducha Charlieho Chaplina****. O* ***to*** *už se přihlásil Steve Martin.* (PCEDT)

As an actor, Charles Lane ***isn't the heir of Charlie Chaplin's spirit****. Steve Martin has already laid his claim to* ***that****.*

[24] According to Mikulová et al. (2005), textual ellipsis also occurrs in grammatical coreference relations (control, dual dependency, reciprocal relations, see Examples 25–27).

[25] The subject is elided in Czech (see description of Czech pro-drop nature in Chapter 1).

4. Nominal groups with a noun head. These are core nouns like *John*, *leg* and nominal groups with governing nouns (see, for example, coreferential relation between *John* and *her son* in Example 12).

Also the reference potential of possessive adjectives like *podnikatelův* [*entrepreneur's*] can be looked upon in the same way as coreference of nouns,[26] see Example 32.

(32) *Tímto faktorem je **podnikatel**, který se snaží o zisk... **Podnikatelova** odměna, zisk, má však svůj původ... v rozbití stacionárního systému.* (PDT)

*This factor is **the entrepreneur**, who is trying to earn a profit... However, the **entrepreneur's** profit is based on... the destruction of the steady system.*

There are certain differences in how noun phrases with different referential potential can enter into textual coreference relations. This issue is addressed in more detail in Section 3.4.

5. Numerals in the position of syntactic nouns. Similarly as noun phrases governed by a noun, numerals in the position of syntactic nouns can take part in textual coreference relations. For example, the numeral *tři* [*three*] is referential in Example 33, but not in Example 34.

(33) *Vybrali **tři** a snědli **je**.* (Mikulová et al., 2005)

They[27] *chose **three** [ones] and ate **them**.*

(34) ***Tři počítače**, které změnily tvář práce s počítači, byly uvedeny na trh v roce 1977.* (PCEDT)

***Three computers** that changed the face of personal computing went on the market in 1977.*

6. Temporal, local and manner pronominal adverbs. Some types of adverbs can be anaphors in coreferential relation, i.e. they can refer to their antecedents in the preceding (or following) context and be substituted by them. These are such anaphoric adverbs as *tam* [*there*], *tehdy* [*then*], *tak* [*so*] (see Example 35) and so on. However, the set of such adverbs in Czech is rather small and it could be defined by a more or less closed list.

(35) *Samozřejmě, že kdyby film obsahoval dialogy, byl by Laneův Umělec nazván **bezdomovcem**. Ale ze stejného důvodu by **tak** říkali i Malému tulákovi.* (PCEDT)

*Of course, if the film contained dialogues, Mr. Lane's Artist would be called **a homeless person**. **So** would the Little Tramp, for that matter.*

[26] In the FGP, they are understood as nouns and respective nouns are reconstructed in the tectogrammatical structure (Mikulová et al., 2005).

[27] The subject is omitted in Czech (see description of Czech pro-drop nature in Chapter 1).

7. Adjectives. Basic adjectives do not refer to discourse entities. However, there are several kinds of adjectives that may have coreferential potential in some contexts. These are, first of all, adjectives which are derived from proper nouns such as *japonský* [*Japanese*] in Example 36 or *pražský* [*Prague,* lit. *Praguian*] in Example 37.

(36) *Podle analytiků projdou v nadcházejícím desetiletí americko-**japonské** vztahy zkouškou, neboť **Japonsko** si uvědomuje svůj nový status ekonomické velmoci regionu.* (PCEDT)

*In the coming decade, analysts say, U.S.-**Japanese** relations will be tested as **Japan** comes to terms with its new status as the region's economic behemoth.*

(37) *Přijel do **Prahy** a **pražská** atmosféra se mu zdála celkem neformální.* (PCEDT)

*He arrived in **Prague** and found the **Prague** atmosphere to be quite casual.*

More than that, in some contexts, adjectives with possessive meaning may be interpreted as referential. In Czech, these are adjectives like *dětský* [*child's, children's, childish, childlike,* etc.] in such cases as *dětská mysl* [*children's mind*].

There is quite a vague border between referring and non-referring adjectives. Even with adjectives having obvious referential potential, it is often not easy to find to which entity (group of entities) they refer.

8. Verbs. Verbs (verbal phrases, clauses, sentences, etc.) have no referential potential and cannot corefer with other verbs. However, verbal expressions can be antecedents of noun phrases in the anaphoric position, as in Example 38.

(38) *Jistotu v tomto směru dávají nejnovější kroky vlády SR, **která se rozhodla zavést již před časem avizovanou desetiprocentní dovozní přirážku na zboží zahraniční provenience**. Na **tento krok** má určité právo.* (PDT)

*In this respect, confidence can be derived from the newest steps of the Slovak government, which decided **to introduce the previously announced 10% tax on goods imported from abroad**. It has the right to make **this step**.*

3.4 Coreference of Nominal Groups with Different Referential Potential

As mentioned above, coreferential relations are most common between nominal groups. However, not all nominal groups are referential in the same way. Similarly, coreferential relations are clear with some types of nominal groups and less clear in other cases.

Unambiguity of coreference relations strongly depends on referential capacity (i.e. how nominal groups of a given kind refer to discourse entities) of noun groups that take part in these relations. There is a number of theoretical research on this topic. For Czech, these are, for example Palek (1968); Palek (1988); Hlavsa (1972); Hlavsa (1975); Adamec (1980), etc. These studies cannot be addressed here in detail, but most authors agree with the distinction between referential (specific and generic) and non-referential (first of all, predicative such as *doctor* in the sentence *Peter is a doctor*) use of noun phrases. Furthermore, textual coreference as the identity of discourse entities in coreferring expressions is not a fully integral phenomenon, some borderline identity relations can be "more identical" than others, thus coreference can be rather observed as a continuum ranging from identity to non-identity. The degree of clarity of coreference relations depends on reference types and on the semantics of anaphoric expressions.

Let us now consider different types of nominal groups with respect to their ability to be "more" or "less" referential and, respectively, coreferential.

1. Coreference of nominal groups with concrete semantics and specific reference
This is mostly the clearest case, cf. obvious coreference relations in the introductory Example 12.

2. Coreference of concrete unspecific nominal groups. The situation is similar with concrete unspecific but not generic entities. For example, in 39, coreferential relation between *some colleagues* and *they* is clear, because such nominal groups become specific once being used anaphorically.

(39) *I will ask* ***some colleagues*** *about it and* ***they*** *will advise me.*

3. Coreference of generic nominal groups. Reference to the type differs from the reference to a concrete object, as it need not refer to all objects of that type. For example, in sentence 40, the word *children* does not refer to all existing children, but to a children prototype (because there are also children, who don't like chocolate). Thus, the question arises, whether references to the same type can be considered to be coreferential. Let us continue Example 40 with the sentence 40a. The sets of children in Example 40 and 40a are not necessarily the same, because there may be a subset of children who like listening to fairy tales but don't eat chocolate, or vice-versa.

(40) *Děti milují čokoládu.*
Children love chocolate.

(40a) *A kromě toho děti také rády poslouchají pohádky.*
Children also like listening to fairy tales.

On the other hand, the repetition of the same expression with generic reference is very important for text cohesion. Similarly, as noun phrases with specific reference, generic expressions can be used anaphorically and in some contexts easily pronominalized (cf. parallel syntactic constructions in Example 41 and 41a, with specific and generic reference of *child/children*, elided or repeated with definite articles or demonstrative pronouns).

(41) *Moje dítě miluje čokoládu. Vždycky chce, abych mu ji koupila.*
My child loves chocolate. He always wants me to buy it for him.

(41a) *Děti milují čokoládu. Proto vždycky chtějí, aby jim ji rodiče kupovali.*
Children love chocolate. They always want parents to buy it for them.

The dividing line between specific and generic use of noun phrases is vague, and sometimes depends on the interpretation. Mostly, both interpretations are possible. Compare the examples below, where in Example 42 we incline to the generic interpretation and in Example 43 to the specific one:

(42) *Pracovníci zahraničních firem působících v České republice často tvrdí, že* ***naši zaměstnanci*** *nedosahují takových kvalit, jaké potřebují... Jsou stesky na nekvalitní výkony* ***našich lidí*** *oprávněné?* (PDT)

Employees of foreign companies based in the Czech Republic often claim, that ***our workers*** *do not have the necessary skills... Is the criticism of the low productivity of* ***our people*** *fair?*

(43) *U* ***detergentu Toto*** *jsme například řešili problém s udržením stálé kvality... Investovali jsme dva miliony korun... a jakost* ***pracího prášku*** *stabilizovali.* (PDT)

For example, with ***the Toto detergent*** *we faced problems with maintaining consistent quality... We invested two million Czech crowns... and stabilized the quality of* ***the detergent****.*

4. Coreference of abstract nouns. Another problematic group for determining reference, and respectively coreference, is the group of abstract nouns. Nouns with abstract meaning make the borderline between referential expression with concrete meaning and predicative parts of speech such as adjective, adverbs and verbs.[28] The basic distinction between abstract and concrete nouns is that concrete nouns refer to material tangible objects (*tree*, *stone*, *paper*, *hair*, etc.), whereas abstract nouns refer to non-material ones (*feeling*, *love*, *imagination*, etc.). Although the distribution of nouns among abstract and concrete is fundamental (see already in Frege, 1892), both groups are quite dynamic, the borderline between them is vague and cannot be unambiguously determined (see different classifications of abstract and concrete nouns in Stepanov, 2004; Arutunova, 1976; Chernejko, 1997, etc.). Looking at theoretical research, the problem of reference potential of abstract nouns seems to be relatively clear: Abstract nouns can evidently corefer, for example, as *his love* in Example 44.

(44) ***He loved her*** *his whole life and* ***his love*** *educated and cultivated him.*

[28] By predicativity of these parts of speech we mean that they do not name an entity, but assign it some qualities.

However, naturally occurring data appear to be much more problematic as for recording coreference relations by abstract nouns. See the following Example 45, where the author speaks about the same feeling (*strach* [*fear*]), but without any anaphoric relation to its previous mention.

(45) *Přiznal, z čeho má **strach**... Všechno nakonec dobře dopadlo, ale tohle dítě zbytečně prožilo půl roku **strachu** a děsivých představ.* (PDT)

lit. *He admitted the origin of **his fear**... In the end, everything turned out well, but the child had to go through half a year of **fear** and horrible thoughts.*

On the contrary, in Example 46 both abstract expressions *originální nápad* [*original idea*] and *nápad* [*idea*] have generic meaning and it is only possible to speak about coreference between the expressions if we understand coreference very generally.

(46) *Asi bych skutečně **originální nápad** v žádosti o grant neuvedl nebo alespoň nesdělil otevřeně. Konkurence ve vědě existuje a stávající systém poskytuje navrhovatelům jen malou ochranu proti zcizení **nápadů**.* (PDT)

*I think I wouldn't reveal my **original ideas** when applying for grant. There is competition in science, and the present system gives the authors very little protection against stealing **ideas**.*

5. Coreference of verbal nouns. Similarly, coreference of verbal nouns is often problematic. On the one hand, they may have a predicative meaning. On the other hand, their predicative interpretation does not prevail in all contexts. Referential potential of verbal nouns is addressed e.g. in Krejdlin and Rachilina (1981) and Mendoza (2004). Mendoza (2004) describes reference of deverbal nouns in the same way as for other noun phrases, with some restrictions to their referential capacity. On the other hand, Krejdlin and Rachilina (1981) single out verbal nouns with specific (Example 47), generic (Example 48) and other types of reference.

(47) *Herečka si jen těžko zvykla na **posun kamery**.*

(Translated from Krejdlin and Rachilina, 1981)

*The actress could not get used to **the camera movement**.*

(48) *Promluvil proti **pronásledování černochů** jako typické formě rasismu.*

(Translated from Krejdlin and Rachilina, 1981)

*He opposed **the persecution of Blacks** as a typical form of rasism.*

If we accept that verbal nouns have the same referential properties as other nouns, we should also accept, that they can take part in coreferential relations in the same way. Thus, if both members of the relation are verbal nouns with a concrete meaning, coreference between them may be considered in the same way as for other concrete nouns, cf. the instances of Czech verbal noun *přiznání* (significantly translated into English as a concrete specific nominal group *the tax return form*) in Example 49.

(49) *Příslušnou rubriku najdete na 2. straně tiskopisu* ***přiznání****. Doklady k odpočtu se k* ***přiznání*** *nepřikládají.* (PDT)

You will find the relevant section on page 2 of ***the tax return form*** (lit. ***confession, declaration***). *The documents are not to be attached to* ***the tax return form****.*

If an anaphoric element is a verbal noun with an abstract meaning, its specific and generic interpretation is possible. In the case when (i) both verbal nouns are specific and refer to a specific situation and their possible arguments are coreferential, or (ii) the anaphoric nominal group refers to a proposition, the relation between them is similar to coreference of specific nominal groups.

Verbal nouns that have generic reference themselves or include generic arguments, corefer similarly to other non-verbal nominal groups with generic reference, see Example 50:

(50) *Rychlé, avšak i bezpečné* ***vypořádání****. Rychlost* ***vypořádání*** *burzovních obchodů... odpovídá potřebám.* (PDT)

Fast, yet safe ***transaction****. The speed of* ***stock-exchange transactions****... corresponds to demands.*

If both verbal nouns have specific reference, but their arguments are not coreferential, or if one verbal noun is generic and the second one has specific reference, these verbal nouns are not considered to be coreferential.

3.5 Coreference Annotation in the PDT

In previous sections, we presented our approach to the phenomenon of coreference. Now, we will describe how this approach is realized in the annotation of Czech textual data in the Prague Dependency Treebank.

Generally, we can say that we attempted to apply our understanding of coreference as close to the theoretical approach described in Sections 3.1–3.4 as possible. This means that we distinguish between grammatical and textual coreference, exophoric and endophoric reference, we take into account different kinds of nominal groups according to their ability to refer, consider coreference by temporal and local adverbs, annotate discourse deixis and so on.

However, annotation of a large-scale corpus demands many separate solutions that should be specified in order to be reasonably compared with other similar annotated corpora. Technical details and phases of the annotation process are presented in Chapter 6. Results for the annotation agreement measurement can be found in Chapter 7. This section addresses annotation solutions, conventions and decisions applied during the annotation of coreference in the PDT. It also describes the decisions made concerning some problematic cases typical for the task of annotating coreference on large-scale corpora.

3.5.1 Scope of annotated expressions

Coreference annotation in the PDT follows the *principle of maximal size of coreferential expressions*. It says that it is always the whole tectogrammatical subtree of the antecedent/anaphor, which is the subject to the annotation. The subtree consists of a governing expression and all its dependents, including subordinate (first of all relative) clauses. So, in Example 51, the whole nominal group *vlády SR, která se rozhodla zavést již před časem avizovanou desetiprocentní dovozní přirážku na zboží zahraniční provenience* [*Slovak government, which decided to implement the previously announced ten-percent import surcharge on goods of foreign origin*] is considered an anaphor. The semantic heads are provided by tectogrammatical representation – this is always a governing node in the corresponding tectogrammatical subtree. In Example 51, the semantic head of the anaphoric nominal group is *vláda* [*government*].

(51) *Nová striktní omezení* ***vlády Slovenské republiky*** *proti českým exportérům... Jistotu v tomto směru dávají nejnovější kroky* ***vlády Slovenské republiky, která se rozhodla zavést již před časem avizovanou desetiprocentní dovozní přirážku na zboží zahraniční provenience****.* (PDT)

The new strict restrictions of ***Slovak government*** *against Czech exporters... In this respect, confidence can be derived from the newest steps of* ***the Slovak government, which decided to introduce the previously announced 10% tax on goods imported from abroad****.*

3.5.2 Embedded nominal groups

We assume that referential properties of nominal groups are not primarily dependent on their syntactic properties. Therefore, it is not essential for postulating coreferential relations, whether they are embedded or not. Thus, apart from the main phrases, we annotate all embedded phrases in the same way as we do for the main ones. So, in Example 52, the nominal group *král s krabicí gumy* [*the king with the box of gum*] has three coreferential links: the whole nominal group, *krabice gumy* [*the box of the gum*] and *guma* [*the gum*].

(52) *I* ***král*** *dostal svou* ***krabici gumy****... Samozřejmě, že novinové zprávy o* ***králi s krabicí gumy*** *byly reklamou k nezaplacení.* (PDT)

The king *also got his* ***box*** *of* ***gum****... Undoubtedly, the articles about the* ***king with the box of gum*** *was an inestimable advertisement.*

The exception for the rule of annotating embedded nominal groups are multiword proper names (*named entities*). Within named entities, we annotate only those embedded nominal groups that are named entities themselves and refer to a different discourse entity. So we annotate coreference for *Česká republika* [*Czech Republic*] inside the phrase *Ústavní soud České republiky* [*the Constitutional Court of the Czech Republic*] but we do not annotate coreference for *výzkum rodiny* [*research on family life*] inside *Oddělení pro výzkum rodiny* [*Department for research on family life*]. We also do not annotate coreference of *republic* within *Czech Republic*, or *John* within *John Smith*, because it refers to the same discourse entity.

3.5.3 Syntactic zeros

On the tectogrammatical layer of the Prague Dependency Treebank, zero arguments are reconstructed using the PDT Valency Lexicon VALLEX (Hajič et al., 2003), that for each autosemantic, valency-capable word unit contains its valency information. According to the detailed classification of ellipses introduced in Mikulová (2011), the PDT uses a rich variety of newly established nodes occupying positions of all kinds of modifications. The classification of these nodes corresponds to the ability of different types of newly established nodes to take part in coreference relations. Here are the newly established tectogrammatical nodes that are subjects to coreference annotation:

The lemma #*Cor*. This lemma is assigned to newly established nodes representing the (usually inexpressible) controllee in control constructions. These nodes are always connected by a grammatical coreference link with its controller, cf. coreference of unexpressed Actor[29] of the verb *pojistit se* [*to insure oneself*] in Example 53, which corefers to the addressee of the governing verb *doporučit* [*to advise*].

(53) ***Čtenářce*** *jsme doporučili ∅ pojistit* ***se*** *u První americko-české pojišťovny.* (PDT)
We advised our ***reader*** *to ∅ insure* ***herself*** *at the First American-Czech Insurance Company.*

The lemma #*Rcp*. This lemma is assigned to newly established nodes representing participants that are left out in the surface form of the sentence in case of reciprocation. There is always a grammatical coreference relationship indicated in the tectogrammatical tree, going from the node with the #*Rcp* t-lemma to the node it is in the reciprocal relation with, cf. the relation between the subject *lovers* and an unexpressed object in the sentence *The lovers kissed #Rcp.PAT*.

The lemma #*PersPron*. This lemma is assigned to nodes representing personal or possessive pronouns. It applies both to newly established nodes and to those present at the surface level. In most cases, nodes with #*PersPron* tectogrammatical lemma are connected with their antecedents by coreference relations (the rare exceptions are mostly generic uses of pronouns used once in the text without further reference). See Example 54 and Figure 3.1.

[29] For the meaning of tectogrammatical functors see Section 6.1.

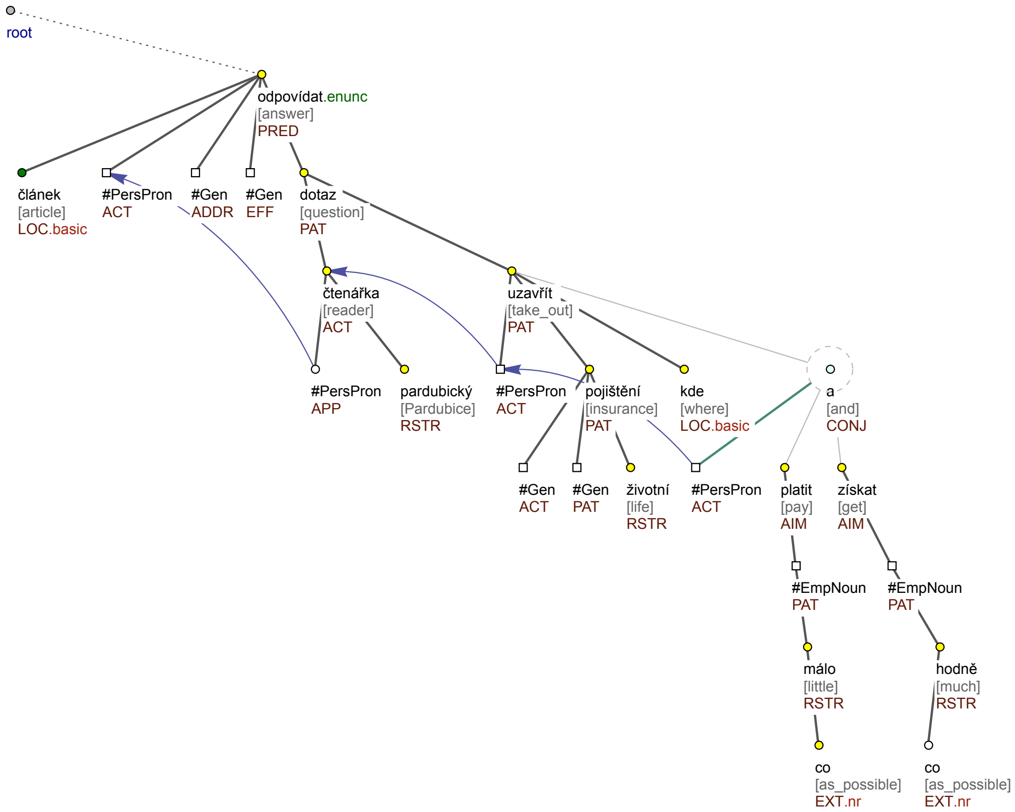


Figure 3.1: Coreference with the reconstructed tectogrammatical node in Example 54

(54) *V článku jsme odpovídali na dotaz **naší pardubické čtenářky**, kde by ∅ měla uzavřít životní pojištění, aby ∅ platila co nejméně a ∅ získala co nejvíce.* (PDT)

lit. *In the article, we answered a question from **our reader from Pardubice**, where ∅ should take out life insurance so that ∅ would pay as little as possible and ∅ get as much as possible.*

*In the article, we answered a question from **our reader from Pardubice**, in which **she** wanted to know where to take out life insurance so that **she** would pay as little as possible and get as much as possible.*

Repetition of a lemma. If it is clear (and possible to identify) which noun has been omitted in the surface structure of the sentence (the case of textual ellipsis), a copy of the node representing the same lexical unit as the omitted element is inserted into the appropriate position, and the coreferential relation with the explicitly expressed antecedent is annotated. So, in Example 55, the noun *pojišťovna* [*the insurance company*] is elided at the end of the sentence. It is reconstructed on tectogrammatical layer and a coreferential relation to an antecedent in the preceding context is annotated (see Figure 3.2).

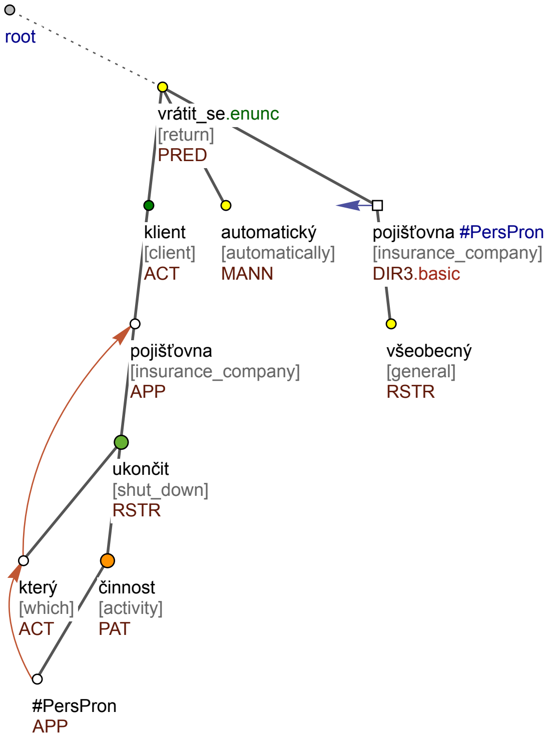


Figure 3.2: Coreference with the reconstructed tectogrammatical node in Example 55

(55) *Klienti pojišťoven, které ukončí svou činnost, se automaticky vrátí k* ***Všeobecné****.*

(PDT)

Clients of insurance companies which shut down will automatically return to the ***General one*** (lit. *General* ∅).

Other newly established nodes are not supposed to be linked by coreference. These are e.g. the tectogrammatical lemmas *#Gen* for a general participant (*Houses are built from bricks*), *#Unsp* for valency modifications with vague (non-specific) semantic content (*U Nováků dobře vaří* [*They cook well at Nováks'*]), *#EmpNoun* for non-expressed nouns governing syntactic adjectives, which are not the case of textual ellipsis (*Přišli* *jen ∅mladší* [lit. *Came only younger* meaning: *Only young people came*]), *#Oblfm* for obligatory adjuncts that are absent at the surface level (*Ta vypadá* [lit. *That*$_{\text{fem}}$ *looks* meaning: *She looks awful/so strange*]) and some other newly established nodes used in comparative constructions.

3.5.4 Non-referring expressions (apposition, predication, verbal complements)

Non-referring expressions such as appositions, verbal complements and nominal groups in predicative position are a special problematic issue in coreference annotation projects. In the PDT, appositions, verbal complements and noun phrases in predicative positions are resolved on the tectogrammatical level in the dependency tree, and are not additionally annotated for coreference. This information can be easily extracted from the tectogrammatical layer. Thus, for appositions, the whole appositive construction serves as an antecedent/anaphor of coreference relations, its parts are connected with a node with a special tectogrammatical functor *APPS* (apposition). The predicative relation is the relation between nominal groups that are sisters in the dependency tree and (except for some special cases) direct daughters of a node with the tectogrammatical functor *PRED* (predicate). For verbal complements, the tectogrammatical functor *COMPL* (complement) is used, the dependency on a noun is additionally represented by means of a special attribute *compl.rf*.

3.5.5 Coordinative constructions and the problem of split antecedents

Coordinative structures and their connection with plural reference are another difficult issue for processing coreference relations. For example, the semantics of plural reference to a coordination like ***John and Mary*** *met.* ***They*** *had not seen each other for a long time* is fairly uncontroversial from a semantic point of view and can be solved satisfactorily by any annotation system (the coordination construction as a whole and its parts separately may be linked by coreference relations). On the other hand, the problem of multiple antecedent for *they* in Example 56 presents a problem for all, no matter if it is a dependency-based or raw-text annotation system. In the PDT, we solve such cases by annotating a bridging relation of the type *set–subset* (see Chapter 4).

(56) ***John*** *visited* ***Ellen****, and* ***they*** *went to the seaside.*

Annotating coreference link for *the Queen* in Example 57 from *Alice in Wonderland* is problematic on the raw text because its modifier *of Hearts* is common for both noun groups, *the King* and *the Queen*. In the PDT, this problem is resolved by a dependency structure. Coordinative elements are represented as direct daughter nodes of a node representing a coordinating connective or operator (see Figure 3.3), shared modifiers are also annotated as direct daughters of coordinating connective or operator (distinguished by values of a special attribute *is_member*). Thus, marking coreferential links between the node of *Queen* in the first sentence and *the Queen* in the second sentence of this example, the modifier *of Hearts* is automatically included in the scope of the antecedent.

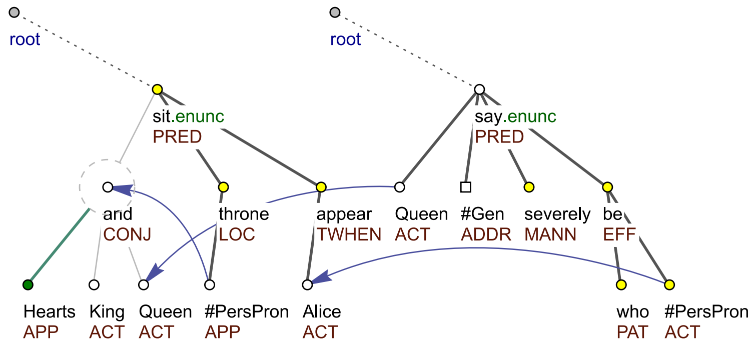


Figure 3.3: Coordinative constructions and split antecedents (Example 57)

(57) ***The King and Queen of Hearts*** *were sitting on their throne when Alice appeared.* ***The Queen*** *said severely "Who is she?"* (Caroll, 1865)

3.5.6 Coreference with specific and generic nominal groups

In the tectogrammatical structure, referring expressions are not classified further into specific and generic ones. Nevertheless, we assume generic nominal groups to have other anaphoric properties in the discourse. Additionally, they result in greater ambiguity and are the cause of lower inter-annotator agreement. Therefore we decided to place them into a special category of coreferential relations.

In the annotation of coreference in the PDT, we distinguish between coreference of nominal groups with specific reference and coreference of nominal groups with generic reference. The information about the type of coreference is obtained from the values *SPEC* (specific) and *GEN* (generic) of the tectogrammatical attribute *informal-type* of the coreference relation.

In some cases, it is hard to define, whether a nominal group has a specific or a generic reference. Mostly, both interpretations are possible. There are no firm rules for an unambiguous assignment of the types in these cases; the type is chosen on the basis of the available context and the annotator's consideration. In ambiguous cases with concrete nouns, the coreference type *SPEC* is preferred.

3.5.7 Discourse deixis

For the reference to verbal phrases, clauses or a sentence, textual coreference links are annotated in the PDT. If a nominal group refers to a segment consisting of more than a single sentence, the special label *segm* is used. However, in this case, the antecedent is not specified. This is a temporary decision and we are planning to specify the scope of larger antecedents in future.

3.5.8 Prepositional phrases

In the PDT, prepositions are hidden in subfunctors and are not represented in tectogrammatical structure. Although the semantic distinction between preposition phrases with the same head and different preposition is very important, it is ignored in the coreference annotation. So, if two nominal groups are coreferential, the coreference relation between them is also marked in the case when they are parts of prepositional phrases which are not coreferential. This is typical for prepositional nominal groups with temporal meaning like *before the war* and *after the war* and for local descriptions like the relation between *za Prahu* [*away from Prague*] – *z Prahy* [*from Prague*] in Example 58.

(58) *Zatím se posunuje stále více* ***za Prahu****... Po dálnici bychom se měli svézt* ***z Prahy*** *až do Českých Budějovic...* (PDT)

So far, people are moving ***away from Prague****... Highways should take us* ***from Prague*** *all the way to České Budějovice...*

3.6 Summary

In this chapter, we have described coreferential relations, how they are interpreted in the Prague approach and how they are annotated in the Prague Dependency Treebank. As we have shown, our understanding of coreference is relatively broad. Apart from pronouns and nominal groups with specific reference, we take into account coreference of some adjectives, local and temporal adverbs, as well as verbal and abstract nouns. Elided expressions reconstructed on the tectogrammatical layer in the PDT made it possible to consider coreferential relations involving syntactic zeros, too. Czech is a language without articles, so it has no formal grammatical means for marking definiteness of nominal groups. For the addressee, as well as for coreference resolution systems, it is quite difficult to distinguish definite anaphoric expressions from indefinite or generic ones. For this reason, both specific and generic nominal groups are annotated for coreference in the PDT. This is also the reason why we focus on coreference, not on anaphoric relations: without a definite article as a formal means for identification of anaphoric expressions, this task will not be fully successful.

4
Bridging Relations

In the previous chapter, we addressed coreference relations, i.e. relations between expressions referring to the same discourse entity. Both coreferential anaphoric and coreferential non-anaphoric relations are highly important for establishing and maintaining textual coherence. However, these are not the only relations between noun phrases that contribute to text cohesion. Take a look at the following Example 59:

(59) *Po babičce nám zůstala* ***starožitná skříňka****.* ***Dvířka*** *se špatně otvírají a skřípou.*
We have inherited ***an antique cupboard*** *from our grandmother.* ***The doors*** *don't open properly and they squeak.*

Here, for the correct interpretation of the noun phrase *the doors*, the addressee accepts the presupposition that the given doors are unique in this context, and the implication that these doors are part of the cupboard mentioned in the preceding sentence.

The definite article used in *the doors* in English and the topical position of *dvířka* in Czech make this inference even stronger. If we assume that *dvířka* [*the doors*] are not part of the cupboard mentioned before, the text would be incoherent.

Such relations are differently defined and classified depending on the approach used. We will characterize them as an inference about two non-coreferential expressions introduced in a text that shows that they are related in some particular way that is not explicitly stated; this relation, however, essentially contributes to text coherence. Thus, generally speaking, these are a kind of coherence-relevant relations between entities which go beyond the notion of coreference.

4.1 Typology of Bridging Relations

In text linguistics, as well as in computational linguistics, a large variety of terms can be found for relations between non-coreferential nominal expressions that influence text coherence. These are, for example *bridging* or *bridging anaphora* introduced in Clark (1975) and used in Asher and Lascarides (1998); Poesio, Vieira and Teufel (1997); Hou, Markert and Strube (2013), etc. The term *inferrables* is used in Prince (1981). To avoid the idea of inference, the term *indirect anaphora* is used in cognitive linguistics, e.g. in Schwarz-Friesel (2007), the relations are sometimes called *associative anaphora*, e.g. in Löbner (1996); Charolles (1999); Miéville (1999) and so on. In this chapter, we will use the term *bridging relations*.

A vague definition of bridging relations includes various bridging scopes in different approaches. For the time being, there is no single generally accepted classification of bridging relations. The basic principle applied in most of existing approaches is that different types of bridging relations are defined on the basis of different kinds of semantic relations between corresponding language expressions.

For example, in Clark (1975), non-identity semantic relations between entities are classified into three main groups: set–subset relation[30] (Example 60), indirect reference by association (Examples 61 and 62) and indirect reference by characterization (Example 63).

(60) *I met* ***two people*** *yesterday.* ***The woman*** *told me a story.* (Clark, 1975)

In the case of reference by association, the bridging anaphor often has as its antecedent some piece of information that is not directly mentioned, but closely associated with the mentioned object. This class represents a wide range of meanings, within which Clark (1975) mentions necessary parts (*room – ceiling*), probable parts (e.g. in Example 61, there is no guarantee that going shopping means walking) and inducible parts (e.g. in Example 62, one has to infer that going shopping included some climbing).

(61) *I went shopping yesterday.* ***The walk*** *did me good.* (Clark, 1975)

(62) *I went shopping yesterday.* ***The climb*** *did me good.* (Clark, 1975)

Indirect reference by characterization primarily describes the situation and its participants, e.g. the relation between the situation of dying and murderer in Example 63:

(63) *John was murdered yesterday.* ***The murderer*** *got away.* (Clark, 1975)

Clark's classification aims to explain contextual boundness of anaphoric expressions. A similar but more detailed classification of non-coreferential semantic relations is presented in Daneš (1979). The author analyses anaphors, contextual boundness of which is given by an explicit expression of the antecedent in the previous context. He operates with the notions of semantic similarity (*sémantická podobnost*) and semantic relatedness (*sémantická souvislost*). In the first case, the semantic structure of both elements of the relation (antecedent and anaphor) has a common denominator in terms of semantic features. These are the relations of inclusion (Example 64), co-hyponymy (e.g. *mother – father* as family members), etc. Semantic relatedness (continuity) may be exemplified by meronymic relations in a broader sense (part–whole and set–subset relations), relations of belonging (a person and his/her clothes), some symptomatic relations (e.g. *fever – illness*) and so on.

[30] Henceforth, we mark general notions of bridging relations in Roman type (e.g. set–subset relation), the relations annotated in the PDT in italics (e.g. *set–subset* relation; not every occurrence of a possible set–subset relation was indeed accepted as a *set–subset* relation in the PDT), without signifying the direction of the relation. Capital letters are used for abbreviated annotation marks in the PDT, which capture the direction of the relation, too (e.g. *SET_SUB, SUB_SET*).

A comparison of Daneš's and Clark's classifications reveals substantial differences. Both the perspective of classification and distribution within each group are different. For example, one of the most stereotypical bridging relations – a set–subset relation – is considered to be a separate group in Clark's classification and a semantic similarity, together with hyponymic-hyperonymic relations in the Daneš classification. Another stereotypical bridging relation – a part–whole relation – belongs to the same category with roles and reasons in Daneš, whereas Clark, again, designates it as a special group. Thus, the approaches seem incomparable, each being a specific and motivating insight into the given problem.

On the other hand, both conceptions are aimed at theoretical research, both are rather descriptive and both try to portray the situation in the languages as deeply as possible. For this reason, the examples given there are not exhaustive and the different types are mostly presented as gradual scales with detailed inner classification.

It should also be noted that usually the term *bridging relation* is used for definite nominal groups (see e.g. Löbner, 1996; Poesio, Vieira and Teufel, 1997). However, the same kind of implicit anaphoric linking is also possible with indefinite or quantifying or even generic nominal groups. For instance, in Example 64, a bridging relation can be observed between the generic nominal group *nový VW Golf* [*the new VW Golf*] (Golf is a type of car made by Volkswagen) and an indefinite nominal group *jedním novým golfem* [*one of the new Golfs*] (one arbitrary car of this category).

(64) ***Nový VW Golf*** *je vybaven motorem o síle 110 kW... Dostali jsme možnost se* ***jedním novým golfem*** *projet.* (PDT)

The new VW Golf *is equipped with an engine power 110 kW... We had an opportunity to ride in* ***one of the new Golfs****.*

Though the research of bridging inferences has concentrated mostly on noun phrases, or, more precisely, on definite descriptions (see e.g. Poesio, Vieira and Teufel, 1997), some bridging relations can also be postulated between events (cf. Asher and Lascarides, 1998 and their Example 65).

(65) *John partied all night yesterday. He's going to get drunk again today.*
(Asher and Lascarides, 1998)

Clark (1975) also presents a comparatively broad class of types of bridging relations which are not restricted to object–type antecedents. Apart from part–whole relations, he mentions "roles" in events, as well as reasons, causes, consequences, and "concurrences" that involve events and states rather than individuals (see example sentences 65a–65e).

(65a) *John was murdered yesterday. The murderer got away.* (role) (Clark, 1975)

(65b) *John fell. What he wanted to do was scare Mary.* (reason) (Clark, 1975)

(65c) *John fell. What he did was trip on a rock.* (cause) (Clark, 1975)

(65d) *John fell. What he did was break his arm.* (consequence) (Clark, 1975)

(65e) *John is a Republican. Mary is slightly daft too.* (concurrence) (Clark, 1975)

4.2 Annotation of Bridging Relations in Corpora

Understanding bridging references is crucial for understanding text as a whole. Together with coreference, discourse structure and topic–focus articulation, this is an important means of text coherence. Thus, resolving bridging relations is a challenge for computational as well as theoretical linguistics.

In the last two decades, a number of corpus studies appeared, addressing bridging relations in written and spoken texts. Among annotations of written texts there are two basic tendencies in the typology of bridging relations. The first tendency is annotating all bridging relations without any further specification of subtypes, such as in Hou, Markert and Strube (2013) or Korzen and Buch-Kromann (2011). These studies operate mostly with definite descriptions in texts (nominal expressions with definite article) and analyze their possible references. The second approach is to specify a set of particular subtypes that are subject to annotation. The number of types ranges between a minimum of three classes (e.g. set–subset, set–element of the set and possession in the GNOME[31] and VENEX[32] corpora) and very detailed classifications in other approaches. For example, in the PAROLE corpus (Gardent, Mahuelian and Kow, 2003), the classification is based on semantic relations of different types, e.g. meronymic relations of part–whole type, event–argument, set–subset, element–attribute, elements of the same situation with a common lexical component (e.g. *flight* – *seat* with a common lexical component *transport facility*) and so on. An interesting solution is applied in the Spanish CESS-ECE corpus (Recasens, Martí and Taulé, 2007), where bridging subtypes are annotated for part–whole, set–member and thematic relations, and the rest of the relations of the bridging type (e.g. general "inference" bridging) are annotated but not further specified with a subtype. A similar solution, but with more bridging subtypes, is annotated in the German SemDok corpus (Bärenfänger et al., 2008).

4.3 Annotation of Bridging Relations in the PDT

We set ourselves two goals during the annotation and classification of anaphoric relations in the Prague Dependency Treebank. On the one hand, we wanted to obtain as much consistently annotated corpus data as possible in order to use it as a train corpus for automatic bridging processing. On the other hand, we were interested

[31] See Poesio (2000).

[32] See Poesio et al. (in prep.).

in creating a detailed close-to-language annotation that would allow us to use the resulting corpus as a basis for further linguistic research of anaphoric relations in Czech.

When choosing the annotation method, we had to give up the idea of annotating all possible bridging relations as it was done e.g. in Hou, Markert and Strube (2013) or in Recasens, Martí and Taulé (2007). It would be too complicated and inconsistent, because Czech lacks the grammatical category of definiteness. There are no articles in Czech which could be used as a formal criterion for identifying definite descriptions as markables, and definite nominal groups may but need not have indicators of definiteness (demonstrative pronouns, positions in sentence, intonation, etc.). For this reason, we decided on an annotation of a set of a few specific types of bridging relations. In the Prague Dependency Treebank, we specify the following six types:

– meronymical relation between a part and a whole with subtypes *PART_WHOLE* and *WHOLE_PART*, see Section 4.3.1,
– the relation between a set and its subsets/elements of the set (with subtypes *SUB_SET* and *SET_SUB*), see Section 4.3.2,
– the relation between an entity and a singular function on this entity (with subtypes *P_FUNCT* and *FUNCT_P*), see Section 4.3.3,
– the relation between coherence-relevant discourse opposites (type *CONTRAST*), see Section 4.3.4,
– non-coreferential explicit anaphoric relation (type *ANAPH*), see Section 4.3.5,
– further underspecified group *REST*, see Section 4.3.6.

It is worth noticing that bridging relations connect not only the individual nominal groups in our approach but the whole coreferential chains (e.g. *Peter – he – the man*). Thus, once postulating a bridging relation between two elements of different coreferential chains, it should not be marked again for coreferential expressions later in the texts.

In case there is a possibility of multiple interpretations of a bridging relations, the annotator had to decide which interpretation is preferable. However, the experiment provided in Nedoluzhko and Mírovský (2013) proved that most cases of interannotator disagreements remain unnoticed by human annotators.

4.3.1 Meronymical relation between a part and a whole

The meronymical relation between a part and a whole is one of the basic bridging relations and has a commonly agreed upon definition (see e.g. Poesio, 2000; Gardent, Mahuelian and Kow, 2003; Recasens, Martí and Taulé, 2007; Bärenfänger et al., 2008). In the PDT, the *part–whole* relation is understood as a relation between inseparable parts and has two directions: The type *PART_WHOLE* is used in cases when the antecedent corresponds to the whole of which the anaphor is a part and *WHOLE_PART* for the opposite order.

Prototypical examples are *room – ceiling*, *hand – finger*, *town – street*, *week – Monday*, etc.

The *part–whole* relation is annotated in the PDT in the following cases:

1. In prototypical cases of inseparable parts, which cannot be understood as subsets (type *room – ceiling*, *hand – finger*), see Example 66:

(66) *Jednotlivá **studia** v apartmánech jsou vybavena **kuchyní**, takže je možná individuální příprava stravy.* (PDT)

__Studio apartments__ are equipped with __kitchens__, so everyone can prepare their own food.

2. With expressions referring to places: states, regions, towns, streets etc. (type *Germany – Bavaria – Munich*, *town – street*, etc.)

3. With references to time spans, as in Example 67:

(67) *Dělal jsem bez přestávky celé **týdny**, často v **noci**.* (PDT)

*I worked nonstop for **weeks**, often at **night**.*

There is a number of borderline cases of the *part–whole* bridging relation. First of all, there is an ambiguous distinction from 'no relation', i.e. when the annotator must decide if the relation is subject to annotation or not. For example, this is the case when a potential expression is not a part of a place, but it is located there (e.g. *Munich – museums*, *galleries* and *rare paintings* in Example 68). In such cases, the bridging relation is not annotated. Also the ambiguities between *set–subset* and *object–function* types are quite frequent.

(68) *V **Mnichově** jsou **muzea** a **galerie se vzácnými obrazy**.* (PDT)

*In **Munich**, there are **museums** and **galleries with rare paintings**.*

4.3.2 The relation between a set and its subsets/elements

The relation between a set and its subsets/elements is understood in a broader sense: It includes non-coreferential relations between nominal groups representing subsets and elements of the set. In our classification, we understand subsets formally: An element of a set is a minimal subset of the given set. Similar to the *part–whole* relation, the *set–subset* relation has two directions – the type *SUB_SET* is used in cases when the antecedent corresponds to a subset or an element of the set of which the anaphor is a set, and *SET_SUB* (see Example 69) is used for the opposite order.
Prototypical examples of *set–subset* relations are *drinks – beer – lemonade – soda*, *butterflies – red ones – white ones*, *seminars – first seminar – last seminar*, etc.

(69) *Pokud tedy zrovna nesedí na svém minikřesle v jednací síni, jsou **poslanci** nuceni pobývat buď ve svých klubech, nebo postávat či posedávat po chodbách. Nelze se pak ani divit, že **část zákonodárců** zvolí příjemnější variantu a odchází úřadovat do suterénní restaurace.* (PDT)

*When they are not sitting directly in their little chairs in the courtroom, **deputies** have to either stay in their deputy clubs, or stand or sit around in the corridors. So one cannot be surprised that **some lawmakers** choose to work in the basement restaurant.*

Annotating set–subset relations is often relatively clear by nouns with a specific reference having a concrete meaning (see Example 69). With generic, abstract and verbal nouns, bridging relations are not always evident. With such nouns, the set–subset relation has a different meaning compared to the relations of specific nouns. The most typical pairs are e.g. "generic expression – a specifying example" (see Example 64 above and "category – subcategory" in Example 70). In the PDT, we consider the set–subset bridging relations with generic, abstract and verbal nouns to be relevant for the text coherence and annotate them as such.

(70) *I když konzervativní Anglie jeho čin odsoudila, guma se zde chytila a Británie se pro **žvýkačku** stala bránou do Evropy. Ještě jeden milník si zaslouží zmínku – zrod **bublinové žvýkačky**.* (PDT)

*Although conservative England criticized his actions, gum caught on here and Britain became the gateway to Europe for **gum**. Another milestone worth mentioning is the birth of **the bubble gum**.*

In some cases, the distinction between *part–whole* and *set–subset* subtypes is quite problematic, so the only reason to choose a specific type of bridging relation is the countability of the corresponding nouns. Ambiguity is more frequent with generic, abstract and verbal nouns, but it can also appear with specifying expressions. At the present stage of annotation, the instruction for annotators in ambiguous cases is to classify bridging relations as *part–whole* only in clear cases of non-separable parts. If some doubts exist, the *set–subset* type should be assigned, as in Example 71.

(71) *Ročně by tedy zaplatila na **pojistném**, včetně **úrazového připojištění**, 4 104 korun.* (PDT)

*She would thus pay 4,104 crowns annually for **insurance**, including **the accidental insurance**.*

In texts with many generic noun phrases, the ambiguity of textual coreference between generic nominal groups and bridging *set–subset* relations is frequent. Mainly, it is caused by different depth of the referential interpretation. For example, in 72, the scope of *the sellers* may be understood both as the same with the scope of *kiosk owners* and as its subset (not all the sellers are also kiosk owners). In accordance with the principle of the preference of coreference relations (see Chapter 3), textual coreference is preferred in such cases.

(72) *V Plzni je **stánkařům** k dispozici tržnice... Prodává se také na náměstí, od **prodejců** vybírá poplatek každé ráno správce tržiště.* (PDT)

*In Pilsen, the market... is available to **kiosk owners**. There is also space on the square, the fees are collected from **the sellers** every morning by the market manager.*

4.3.3 The relation between an entity and its singular function

The bridging relation *object–function* is annotated between two entities when one entity has a singular function on another entity. Similar to *part–whole* and *set–subset* relations, the bridging relation *object–function* has two uses: the type *FUNCT_P* is used for the case when the antecedent corresponds to a function on the anaphor which is in the anaphoric position, and *P_FUNCT* for the opposite order.

Prototypical examples of *object–function* relations are *trainer – team*, *prime minister – government*, *company – director*, *event – organizer* and so on.

In Example 73, *the state* (in this case, Czech Republic) has only one operating government.

(73) *Na přímou podporu podnikání vydá letos **stát** přibližně 1,8 procenta hrubého domácího produktu. Tuto skutečnost jednoznačně konstatuje ministr hospodářství Karel Dyba v analýze, kterou předložil **vládě**.* (PDT)

***The state** will give about 1.8 percent of the gross domestic product to directly support business this year. This fact is clearly stated by Economy Minister Karel Dyba in his analysis which he presented to **the government**.*

The distinction between *set–subset* and *object–function* types is defined on the basis of singularity of the given function. For this reason, the relation between *minister* and *government* is marked as *set–subset*, while the relation *prime minister – government* is annotated as *object–function*.

In some cases it is hard to decide if the relation is still coherence-important and should be annotated as bridging or it should be omitted. If the case is ambiguous, it is up to the annotator to decide which interpretation is involved. The recommendation is rather not to mark clearly ambiguous cases.

4.3.4 The relation between coherence-relevant discourse opposites

The *contrast* bridging relation is annotated between nominal groups standing in the relation of discourse opposites. Unlike the relations mentioned above, this relation has only one direction. The relation is marked on the basis of the context, thus it is hard to produce prototypical examples.

In Example 74, the nominal group *the fortunes of the Czech Republic* is put into contrast with *the fortunes of Slovakia*:

(74) *Dnes, po rozdělení Československa, je jasné, že* ***osud České republiky*** *bude stále více spojený s Německem a přes něj s Evropskou unií a* ***osud Slovenska*** *s Ruskem.* (PDT)

Nowadays, after the split of Czechoslovakia, it is clear that ***the fortunes of the Czech Republic*** *will be increasingly tied to Germany and thus to the European Union, while* ***the fortunes of Slovakia*** *will be tied more to Russia.*

The *contrast* relation is not a bridging relation in the restricted sense – it could rather be labelled as a rhetorical relation between nominal groups. However, this kind of semantic dependence has a similar influence on text cohesion as bridging relations.

Moreover, contrast annotated on the level of bridging relations supplements other types of contrast in text annotated in the PDT. Contrastive contextually bound nodes are captured with the topic–focus articulation annotation (see Chapter 5). The relations of *confrontation*, *opposition* and *pragmatic contrast* are annotated on the discourse level (see Chapter 2). The annotation of contrasts on the topic–focus articulation level captures this phenomenon on the level of nodes (individual expression). Contrasts within discourse annotation concern propositions (clauses, sentences and larger textual segments). Contrast as a bridging relation sets the relationship between nominal and prepositional groups.

4.3.5 Non-coreferential explicit anaphoric relation

In a non-coreferential anaphoric relation where the anaphor is marked with an explicit anaphoric marker (demonstrative pronoun or adjective, contextual boundness represented by the word order, etc.), the special bridging relation of the type *non-coreferential anaphora* is annotated. It has one direction, it always refers back to the antecedent.

The bridging relation *non-coreferential anaphora* is marked in the following cases:

1. Metalinguistic references, i.e. references to an antecedent expression, not to an extralinguistic object. This can be illustrated with Example 75, where *re-education* is not coreferential with *the term re-education* but anaphorically refers to it.

(75) ***Termín převýchova*** *znám pouze z nacistického a komunistického slovníku. Na* ***převýchovu*** *se, pokud vím, posílali ti, kteří měli podle těchto zrůdných režimů nevhodný původ.* (PDT)

I know ***the term re-education*** *only from Nazi and Communist vocabulary. As far as I know, the people who were sent for* ***re-education****, were considered to have an unacceptable origin according to these monstrous regimes.*

Metalinguistic references are also common in contexts with attribution (combination of the author's and direct speech), as in Example 76. Here, the reference implied by *rainbow* in the book that the priest was reading is not identical with the reference of *the word, where he stopped*. In the first case, *rainbow* most probably refers to a natural occurrence, the second nominal group *the word, where he stopped* refers to the word in the book.

(76) *„**Duha**?" Kněz přiložil prst k **tomu slovu, kde skončil**.* (PDT)

*"**A rainbow**?" The priest put his finger on **the word, where he stopped**.*

2. Anaphoric reference to the time when the antecedent situation takes place (Example 77).

(77) ***Rozbití Varšavské smlouvy** bylo jako odseknutí údů od těla. Od **té doby** se toho mnoho neudělalo.* (PDT)

***The disintegration of the Warsaw Pact** was like cutting limbs off from the body. Since **that time**, there was not much that was done.*

3. Anaphoric reference to an object which has some similar characteristics to its antecedent. Usually, complements like *takový* [*such*], *podobný* [*similar*], *stejný* [*the same*], etc. are used with the noun phrase in the anaphoric position.

(78) *Nic nenasvědčuje tomu, že by **parlamentní budova měla sloužit jiným než parlamentním účelům**. Přesto se **takové názory** ozývají.* (PDT)

*There is no indication that **the parliamentary building could serve other purposes than parliamentary ones**. However, one sometimes hears **such opinions**.*

This last group may be problematic in cases, where instead of anaphoric expressions of similarity (e.g. *such* and *similar*), anaphoric expressions of difference (e.g. *other*) are used. In this case, the PDT conventions say that the bridging relation *contrast* should be marked.

4.3.6 Further underspecified bridging relations

The *rest* bridging relation is annotated in cases when expressions are connected by a bridging relation which is not included in any of the groups above. This type is used for capturing potential candidates for a new group of bridging relations.

In the present annotation of bridging relations in the PDT, the *rest* group is restricted to the following types:

- the relation "location–resident" by place names, e.g. *Berlin – Berliner*, and common nouns (*state – population*);
- relations between relatives (*mother – son*);
- the relation "author–his work" (*J.R.R. Tolkien – The Hobbit*);

- the relation “event–argument” (*research* – *researcher*);
- the relation “object–typical instrument” (*woodcutter* – *axe*).

Bridging relations that do not fit this description are not annotated as such in the PDT.

4.4 Discussion and Summary

The PDT-style conception of capturing bridging relations in Czech designates a set of specific types that are subject to annotation. Of course, these types do not cover all bridging relations that occur in texts. Many relations remain neglected. We will now discuss, which types of bridging relations remain un-annotated in the PDT and why.

First, rather than syntactic, our annotation of bridging relations has a semantic (sometimes even pragmatic) nature. It means that **bridging relations are not annotated, if they are already captured by the syntactic structure of the tectogrammatical layer**. In the PDT, there are some tectogrammatical functors that include corresponding meaning in their semantics. For example, the tectogrammatical functor *AUTH* (author) that is assigned to nouns that denote the author of an artefact. For this reason, in the sentence item *Tolkien's.*AUTH *Hobbit*, there is no need to annotate the bridging relation between *Tolkien* and *Hobbit*, it is deducible from the tectogrammatical tree. Besides, bridging relations are not annotated by direct dependencies of nouns with tectogrammatical functors *APP* (appurtenance) in *člen týmu* [*member of the team.*APP], *MAT* (material, partitive) in *polovina lidí* [*half of the people.*MAT] and *PAT* (patient) in *obyvatel obce* [*resident of the village.*PAT]. In other words, we did not annotate so-called bridging relations of genitive constructions within a single clause. However, these relations are subject to annotation when there is no direct syntactic dependency between the nodes (e.g. when *village* and *resident of the village* are in different clauses or sentences).

Bridging relation *part–whole* is not annotated by direct dependencies with the tectogrammatical attribute *ACMP* (accompaniment) and bridging *contrast* is neglected with the attributes *ADVS* (adversative) and *CONFR* (confrontation) as these attributes already include the corresponding meanings in their semantics.

Second, having included meronymy in the set of annotated types of bridging relations, we did not include the relation of **co-hyponymy**. Thus, for example, the relation between *body* and *hands* will be annotated as bridging relation of the type *part–whole*, but not the relation between *hands* and *legs*. Making such a decision deprives us of a relatively large number of relations that may be relevant for text coherence. On the other hand, an experiment with annotating co-hyponymy has shown that it brings down the inter-annotator agreement dramatically, as meronymic relations often come into conflict with co-hyponymic ones, making the networks of the relations much more complicated.

Third, our annotation does not mark all kinds of **implicit semantic and pragmatic inferences** such as bridging reference of *the park* to previous context in Clark's Example 79:

(79) *John went walking at noon.* ***The park*** *was beautiful.* (Clark, 1975)

The fact is that bridging relations are based on inferences that people make when reading and interpreting the text and these inferences depend on many factors. Most factors, e.g. situation of speech, relations between speakers and their common knowledge, level of education, gender, age, etc. have extralinguistic nature. It appears to be a serious challenge in many cases for a human annotator to distinguish between semantic and pragmatic knowledge, especially when considering the relations between full autosemantic noun groups. There is a scale ranging from nominal groups which are uniquely interpretable by means of world knowledge to those which depend on a previous anchor. Nevertheless, many real examples remain in between. When trying not to capture the relations based on world-knowledge and not concerning language knowledge at all (for example, not annotating non-obvious relations between named entities designating persons), we still annotate the relation *part–whole* between *Prague* and *Czech Republic*, as it seems to be a kind of "common world-knowledge," since it can be found in WordNet-like databases and so on. "Common world-knowledge" is hard to define, so we could not expect very high agreement on the interpretation of these inferences. However, the extended linguistic and pragmatic analysis of text relations interpretation could help us understand human nature in much more detail.

We can conclude that the classification of bridging relations as well as choosing the scope of annotation strongly depend on the research topic. We believe that the approach described in this section will help us work with the chosen material to develop further theoretical and practical research.

5

Topic–Focus Articulation

5.1 What Is Topic–Focus Articulation

One way to look at discourse is to view it as a sequence of utterances, taking into account the so-called *information structure of the sentence* (*topic–focus articulation*). This aspect of sentence structure is a good "bridge" towards a study of (at least one aspect of) the dynamic development of discourse. This, of course, is not a new idea: To our knowledge, its first comprehensive treatment, though taken from a psychological rather than linguistic perspective, was provided by Weil (1844). According to Weil (1978, p. 11), "Words are the signs of ideas; to treat of the order of words is, then, in a measure, to treat of the order of ideas." Weil recognized two types of "movement of ideas," namely *marche parallèle* and *progression*: "If the initial notion is related to the united notion of the preceding sentence, the march of the two sentences is to some extent parallel; if it is related to the goal of the sentence which precedes, there is a progression in the march of the discourse" (ibidem, p. 41). It should not be overlooked that Weil (ibidem, p. 45) also noticed the possibility of a reverse order which he calls 'pathetic': "When the imagination is vividly impressed, or when the sensibilities of the soul are deeply stirred, the speaker enters into the matter of his discourse at the goal."

In more modern terms, one can say that two adjacent utterances may either be linked by their *topics* or the topic[33] of one utterance may be linked to the *focus* of the preceding one (see the two basic types of thematic progressions in Daneš, 1974).

The readers or hearers of a text are accustomed of being informed from a particular perspective. They expect to receive a certain anchor, i.e. to start with what they have already known and on the basis of this "old" knowledge they accept "new" concepts or new relations among previously mentioned elements. These new concepts or new relations then fit into the previous (con)text and become known. And again, through the information that was just obtained, people can accept more new information. The same principle is usually reflected in the build-up of a text and on lower layer, in the formulation of individual sentences. In this way, topic–focus articulation performs the communicative function of the text.

[33] In different approaches to this domain of study different terminology is used: topic – focus, theme – rheme, background – focus, etc. The underlying ideas are very close to each other, though there are, of course, differences in their interpretation.

5.2 The Importance of Topic–Focus Articulation – Language Comic and Misinterpretation

Besides the above mentioned communicative function, topic–focus articulation is also a language phenomenon that significantly affects the sentence semantics, cf. Example 80.

(80) *Entry with dogs on leash only.*

The sentence in Example 80 can be interpreted in two ways: (i) the entry of dogs is allowed only if they are on a leash, or (ii) the entry is allowed only if you have a dog (on a leash). The intonation center is put on the word *leash* in both cases: *Entry with dogs on LEASH*[34] *only.* The two interpretations vary in the scope of the focus particle *only* (called *focalizer* or *rhematizer*).[35] In the first case, the focus particle *only* concerns the participant *on leash* – while in the second case, *only* pertains to the whole prepositional group *with dogs on leash*.

The misinterpretation of the topic and focus of a sentence may cause misunderstandings between the speaker and the addressee and may also be a source of language comic, see e.g. Example 81.

(81) *Why do we dress baby girls in pink and baby boys in blue? Because they do not know how to dress themselves.*

In the most common interpretation of the sentence, the pronoun *we* stays in the background of our attention; the emphasis is put on the colors of girls' and boys' clothing. However, the answer deals with the pronoun *we* as if it were emphasized: It says why the baby girls and boys are dressed exactly by us (not why they are dressed in pink and in blue as we would probably expect). It should be noted that the position of the intonation center again plays an important role here. Both examples illustrate the importance of the distinction between the information the addressees understand as the topic of the sentence, and the information newly introduced and non-identifiable.

In this chapter, we first describe the theoretical basis and fundamental notions of the theory of topic–focus articulation that we subscribe to, such as *contextual boundness*, *communicative dynamism* and *topic and focus*. In the second part of this chapter, we outline how topic–focus articulation is captured in the Prague Dependency Treebank.

5.3 The Theoretical Basis

The original formulations of what is now more generally referred to as the *information structure of the sentence* were based on a dichotomy, be it a distinction between

[34] The intonation center is henceforth marked in capitals.

[35] For the interpretation of rhematizer, see Hajičová (1995). A detailed analysis of this category based on the PDT material is given by Štěpánková (2014).

psychological subject and *psychological predicate*, *theme–rheme*, *topic–comment*, *topic–focus*, *presupposition* and *focus*, *given* and *new information* etc. In structural linguistics, the pioneer of the study of these topics was Mathesius, who refers to Weil (1844) quoted above, and to linguists around Zeitschrift für Völkerpsychologie, von der Gabelentz (1868), Paul (1886), and esp. Wegener (1885), though criticizing their terms *psychological subject* and *psychological predicate* (Mathesius, 1907). Mathesius himself calls this articulation by the Czech term *aktuální členění* (literally translated as "actual articulation") because it is determined (guided) by the "actual," that is "topical" situation of the speaker and concerns the way, in which the sentence is incorporated into the factual relation to the situation from which it originated (Mathesius, 1939). Mathesius distinguishes between *východiště výpovědi* (initial starting point of the utterance, its basis), which he specifies as "what is known or at least evident in the given situation and from where the speaker starts" on the one hand and *jádro výpovědi* (nucleus of the utterance), that is "what the speaker utters about with respect to the starting point of the utterance." Mathesius prefers the above specification rather than using *known* and *unknown*. However, already in Mathesius' writings we see a certain inclination to recognize a more articulated scale rather than a mere dichotomy, when he says that the starting point may contain more than a single element so that it is possible to speak about the center of the starting point and the accompanying elements which "lead from the center to the nucleus." Referring to the position of the sentence predicate, Mathesius writes that the predicate is a part of the nucleus but on its edge rather than in its center and represents a transition between the two parts of the utterance.

Mathesius' observations inspired the fundamental work of Firbas and his team. As Mathesius' original Czech term *aktuální členění větné* is not directly translatable into English and apparently inspired by Mathesius' use (Mathesius, 1929) of the German term *Satzperspektive* Firbas used the term *functional sentence perspective* (FSP). Very early in the development of the FSP approach, the binary articulation into *theme* and *rheme* was complemented – also in line with Mathesius' ideas mentioned above – by a more structured approach introducing the notions of *transition* and even a more scalar notion of *communicative dynamism* (CD). From this point of view, theme was specified by Firbas (1964) as being constituted by an element or elements carrying the lowest degree(s) of communicative dynamism within a sentence (which was later modified by Firbas (1992) in the sense that theme need not be implemented in every sentence, while in every sentence there must be *rheme proper* and *transition proper*). The concept of communicative dynamism was characterized by Firbas (1971) as a hierarchy of degrees carried by a linguistic element of the sentence, i.e. "the extent to which the element contributes towards the development of communication." The basic distribution of communicative dynamism would then reflect what Weil (1844) called the "movement of the mind."

Almost in parallel with FSP, but also partly as a reaction to it, Sgall and his collaborators in Prague developed the theory of *topic–focus articulation* (TFA) (see e.g. Sgall, 1967b; Sgall, Hajičová and Benešová, 1973; Sgall, Hajičová and Buráňová, 1980; Sgall, Hajičová and Panevová, 1986; Hajičová, Partee and Sgall, 1998). The theory of topic–focus articulation is an integral part of the formal model of Functional Generative Description of language, namely of the representation of sentences on the underlying (tectogrammatical) sentence structure, see Chapter 1 and Chapter 6. These tectogrammatical representations are viewed as dependency trees, with the main verb being the root of the tree. Every node of the tree carries – in addition to other characteristics such as the type of dependency – an index of contextual boundness: a node can be either contextual bound or non-bound. This feature, however, does not necessarily mean that the entity is known from the previous context or new but rather how it is structured as for the information structure of the sentence.

With the help of the bound/non-bound primary opposition, the distinction between the topic and the focus of the sentence can be defined depending on the status of the main verb (i.e. the root) of the sentence. If the verb is contextually bound then the verb and all the nodes depending (immediately or not) on the verb constitute the topic, the rest of the sentence belongs to its focus; if the verb is contextually non-bound, then the verb and all the nodes depending on it to the right constitute the focus, while the rest of the sentence belongs to its topic (see the definition of topic and focus in Sgall, 1979).

The left-to-right dimension of the tree serves as the basis for the specification of the scale of communicative dynamism: Communicative dynamism is specified as the deep word order, with the least dynamic element standing in the leftmost position and the most dynamic element (the focus proper of the sentence) being the rightmost element of the dependency tree.

In spoken language, the most important means of expressing the difference in topic–focus articulation is the sentence prosody including the placement of the intonation center; in our more recent work with spoken language corpora, the characteristics of the curve were considered as a marker of a *contrastive topic* (Veselá, Peterek and Hajičová, 2003).

Currently, the phenomenon of topic–focus articulation is included essentially in most formal (and empirical) language descriptions under different names, such as *information structure* (the term used by a number of authors, e.g. by Steedman, 1991 or Lambrecht, 1996); see also the treatment of *communicative structure* in the Meaning–Text Theory as developed by Mel'čuk (1981).

In our analysis, we use the Functional Generative Description as the main theoretical basis for our linguistic approach and also as the basis for annotating topic–focus articulation in the Prague Dependency Treebank which is a fundamental language data source for the research described; we also utilize the term *topic–focus articulation*.[36]

[36] For comparison of the FGD approach with the further approaches to topic–focus articulation, see Hajičová (1972); Sgall, Hajičová and Benešová (1973); Sgall (1975); Hajičová, Partee and Sgall (1998) or Hajičová (2012).

5.4 Basic Terms of Topic–Focus Articulation

The description of topic–focus articulation is based on three main features: (i) *contextual boundness*; (ii) *communicative dynamism* and (iii) sentence division into *topic* and *focus*. Topic and focus are defined on the basis of the first two characteristics. Therefore, we introduce the contextual boundness and communicative dynamism phenomena first and then describe the conception of topic and focus within Functional Generative Description approach.

5.4.1 Context and contextual boundness

Sentences in a coherent text are interconnected by various types of relationships (explicitly marked or implicitly present) – the relationship of contextual boundness between sentence items and the context is one of them. The *context* can be provided by the previous sentences (i.e. by the previous text or texts) or by the broader setting of situation in which the text is created or perceived. The *situational context* is not fixed and its setting can influence the text perception (e.g. Shakespeare's dramas were understood differently in 17th century than now because the situational context has changed). The situational context includes any shared or generally known information, which may be determined by the immediate situation or longer experience, senses, culture or other factors.

Depending on the context, we can decide for every sentence item (that is relevant for topic–focus articulation) whether it is *contextually bound* or *non-bound*. In the Functional Generative Description, the contextual boundness is a property of an element of the sentence (expressed or absent in the surface sentence structure) which determines whether the author uses the sentence element as given (for the recipient), i.e. uniquely determined by the context, see Hajičová, Partee and Sgall (1998). It means that contextually bound sentence items are deducible from the broader context, see Example 82.[37]

(82) (***Jane** is my best friend.*) ***She** is very NICE.*

The pronoun *she* is contextually bound because it is deducible from the previous context. On the contrary, all other sentence items are contextually non-bound in this case because they bring information that cannot be deduced from the (previous) context.

The relationship of contextual boundness may seem similar to coreferential and anaphoric relations. Nevertheless, they do not necessarily coincide since they describe data from different points of view, cf. Chapter 13. In Example 83, items *her* and *him* have a coreferential relation to some previous sentence items. However,

[37] The sentence in parentheses denotes the context, be it immediately preceding or distant, in which the example sentence is supposed to be uttered.

they are contextually non-bound, as they present the items from the context in a new, indeducible relation.

(83) (*For some Catholics, **Mary**$_1$ is more important than **Christ**$_2$.*) *They go to **HER**$_1$ not to **HIM**$_2$.* (Perspective Digest)

Contrast and contextual boundness

A common way of how information can be formulated is to express it in *contrast* with the known context. This contrastivity is reflected also in the topic–focus articulation structure. Namely, contextually bound sentence elements can stand in contrast as in Example 84:

(84) (*We have two children.*) ***John***$_c$ *is the YOUNGER, **Mary**$_c$ is the OLDER.*[38]

In this example, *John* and *Mary* are presented by the author as contextually bound – they were introduced in the first sentence and now they are referred to as a starting point for the flow of the text in which information about their age is presented. On the other hand, they are presented in contrast to each other, with the background formed by the word *children*.

In the Functional Generative Description, this case is discerned as a special subtype of contextual boundness – the *contrastive contextual boundness*. Its delimitation is broad: Whereas in Example 84, the contrasting elements are mentioned explicitly, there may occur structures in which the contrastivity is implicit, resulting from remoted text segments or world knowledge. The item chosen as a starting point of the sentence may be a part of a set of possible starting items in the given context, cf. Example 85.

(85) (*The weather is nice.*) ***John***$_c$ *is playing in the GARDEN.*

Here, the second sentence could have been started from more items deducible from the situation like temperature, the speaker, the day, the whole family present in the situation. However, the speaker decided to choose specifically *John* to start his utterance, in contrast to the other items of the set of possible alternatives.

There are typical ways to formally express the feature of *contrastivity* for contextually bound items. One of them is *contrastive stress*, as in Examples 84 and 85. In Czech, specific (long) forms of pronouns are used to express contrastivity, while non-contrastive contextually bound pronouns have short forms, cf. Examples 86 and 87 with a long stressed contrastive form *tebe* [*you*] and a short clitic non-contrastive form *tě* [*you*].

[38] Here and in further examples, contrastive contextually bound items are labelled with *c*, non-contrastive contextually bound items bear a mark *t*, the contextually non-bound nodes are marked as *f*.

(86) ***Tebe***$_{\mathrm{c}}$ *já NEZNÁM.*
lit. ***You***$_{\mathrm{c}}$ *I DO_NOT_KNOW.*
Concerning you$_{\mathrm{c}}$*, I DO NOT KNOW you.*

(87) *Já* ***tě***$_{\mathrm{t}}$ *NEZNÁM.*
lit. *I* ***you***$_{\mathrm{t}}$ *DO_NOT_KNOW.*
I DO NOT KNOW ***you***$_{\mathrm{t}}$*.*

So far we have dealt with contrastivity for contextually bound items. For the contextually non-bound sentence members, such a distinction is not considered to be relevant, because as a matter of fact, the newly presented items always concern a choice of alternatives and to some extent stand in contrast to the previous context. Unlike contextual boundness, contextual non-boundness has no special formal means for discerning the feature of contrastivity.

To sum up, the theory discerns two basic categories: *contextual non-boundness*, *contextual boundness* and a subcategory of the *contrastive contextual boundness*.

A distribution of sentence items with various values of contextual boundness is presented in Example 88.

(88) *Across the river*$_{\mathrm{c}}$ *Magda*$_{\mathrm{t}}$ *and Kovarik*$_{\mathrm{t}}$ *could now*$_{\mathrm{t}}$ *see*$_{\mathrm{f}}$ *a FIRE*$_{\mathrm{f}}$ *with two*$_{\mathrm{f}}$ *figures*$_{\mathrm{f}}$ *beside it*$_{\mathrm{t}}$*. When they*$_{\mathrm{t}}$ *moved*$_{\mathrm{f}}$ *closer*$_{\mathrm{f}}$*, they*$_{\mathrm{t}}$ *could make*$_{\mathrm{f}}$ *out two*$_{\mathrm{f}}$ *white*$_{\mathrm{f}}$ *HORSES*$_{\mathrm{f}}$ *against the background*$_{\mathrm{f}}$ *of the dark*$_{\mathrm{f}}$ *bushes*$_{\mathrm{f}}$*. Then*$_{\mathrm{t}}$ *he*$_{\mathrm{t}}$ [*Kovarik*] *RECOGNIZED*$_{\mathrm{f}}$ *them*$_{\mathrm{t}}$*.* (Škvorecký, 1986)

We can observe that mainly the temporal and circumstantial adjuncts in the role of scene setting (e.g. *beside it*, *now*) and subjects presented as given (e.g. *Magda* and *Kovarik*, *they*) are contextually bound. On the contrary, most of predicates (e.g. *could see*, *could make out*, *recognize*) are contextually non-bound because they are not deducible from the context. Contrastive contextually bound sentence items are rather rare in authentic texts. In Example 88 there is only one item marked as contrastive contextually bound sentence element, namely the local setting *across the river*. The location is given by the broader context of the situation but it offers a choice of one alternative out of several others (*on this side of the river*, *at distance*, …) given within that context.

At the same time, we can see that contextually bound sentence items can be modified also by contextually non-bound sentence elements (e.g. *two figures beside it*) and on the contrary, contextually non-bound sentence items can be modified by dependent contextually bound elements, see Figure 5.3.

5.4.2 Communicative dynamism

When observing the sentence and its contextually bound and non-bound parts, we can see that the individual sentence items mutually differ in degrees of their relative importance. Firbas (1971) characterized this phenomenon as *communicative dynamism* and postulated the concept of information hierarchy in the sentence. Firbas likened communicative dynamism to information flow. He claimed that the degree of communicative dynamism is specified as relative importance with which the given element contributes to the development of communication, i.e. to what extent the sentence element moves the communication forward.

The Functional Generative Description took over this concept and applied it in formal description. According to Hajičová, Partee and Sgall (1998), communicative dynamism is a property of a sentence element that reflects its relative degree of communicative importance attributed to it by the author – compared with other sentence elements in the sentence; contextually non-bound sentence elements are considered to be more dynamic than sentence elements contextually bound (be they non-contrastive or contrastive).

Communicative dynamism is not seen as a dichotomy but as a scale with more degrees. Such a scale is reflected in the so-called *deep word order*. Deep word order describes the organization of elements in a sentence structure according to their increasing communicative dynamism. In some cases, deep word order can be directly related to the *surface word order*,[39] see Example 89 from the text about Rusalka from Chapter 1.

(89) *He*$_{\mathrm{t}}$ *looked*$_{\mathrm{f}}$ *at MAGDA*$_{\mathrm{f}}$. (Škvorecký, 1986)

In Example 89, there is one contextually bound item (*he*) and two contextually non-bound items (*to look, Magda*). The contextually bound item carries the lowest degree of communicative dynamism, i.e. the lowest relative degree of importance, and it is followed by contextually non-bound items that carry a higher degree of communicative dynamism. At the same time, the predicate (*to look*) carries a lower degree of communicative dynamism than the element *Magda*, despite the fact that both of them are contextually non-bound.

Empirical investigations of topic–focus articulation in Czech have indicated that the individual values of communicative dynamism are connected with contextual boundness. However, it is supposed that the individual values of communicative dynamism function differently among contextually bound sentence items in comparison with contextually non-bound items (directly dependent on their governing verb). The order of contextually bound modifications directly depending on the verb is determined in the scale of communicative dynamism by the choice of the author and it may be affected by various factors – the language factors (e.g. Actor may be

[39] The *surface word order* is the ordering of sentence elements in the surface structure, i.e. the word order in sentences realized in real texts (for more details, see Rysová and Mírovský, 2014a). The difference between the deep and surface word order occurs more frequently in languages with a grammatically fixed word order (such as English), while with languages such as Czech the surface word order is typically governed by topic–focus articulation and as such corresponds to the deep word order. For more details see Section 5.6.2.

chosen as the least dynamic item more easily than other participants), the situation factors (e.g. whether the entity was mentioned in the immediately preceding context or whether it is not really activated in the consciousness of the author and addressee) or by factors related to the text composition (e.g. use of contrast).

On the other hand, the contextually non-bound verb modifications directly depending on the verb are supposed to follow the so-called *systemic ordering* (Sgall, Hajičová and Buráňová, 1980, see also Zikánová, 2006; Rysová, 2011; Rysová, 2014a), i.e. a scale of communicative dynamism for contextually non-bound sentence items directly dependent on their governing verb. Systemic ordering presumes e.g. that contextually non-bound Patient carries a higher degree of communicative dynamism than e.g. contextually non-bound Temporal modification in English sentences. The existence of systemic ordering in languages is considered to be language independent but the individual degrees of it are language specific (i.e. systemic ordering in Czech is different than systemic ordering in English).

In English, systemic ordering is only rarely reflected in the surface word order. However, in Czech we can study its systemic ordering particularly from the surface word order. In most cases, the contextually non-bound sentence items (directly dependent on their governing verb) also follow the systemic ordering in surface word order because Czech is a language with free word order and its surface word order is affected by communicative dynamism to a large extent (unlike English).

While in English e.g. the order of the members carrying the highest degree of communicative dynamism is mostly grammatically fixed, in Czech they are usually placed at the very end of the sentence, cf. Example 90 from Chapter 1.

(90) *Potom*$_{\mathrm{t}}$ [*on*$_{\mathrm{t}}$] *je*$_{\mathrm{t}}$ *POZNAL*$_{\mathrm{f}}$. (Škvorecký, 1991)
lit. *Then*$_{\mathrm{t}}$ [*he*$_{\mathrm{t}}$] *them*$_{\mathrm{t}}$ *RECOGNIZED*$_{\mathrm{f}}$.
Then$_{\mathrm{t}}$ *he*$_{\mathrm{t}}$ *RECOGNIZED*$_{\mathrm{f}}$ *them*$_{\mathrm{t}}$. (Škvorecký, 1986)

In Example 90, the most dynamic element is *poznal* [*recognized*]. All other sentence items carry a lower degree of communicative dynamism. In Czech, this fact is captured also in surface word order – the most dynamic sentence element is placed in the last position whereas the object *je* [*them*] stands before the predicate. On the contrary, in English, the last item is the word form *them* that is contextually bound and therefore also less dynamic than the contextually non-bound predicate *recognized*. On the basis of this example, we can see that Czech surface word order is much more influenced by communicative dynamism than the word order in English. In English, surface word order is affected more by grammatical factors than by topic–focus articulation.

5.4.3 Topic and focus

On the basis of the previously described phenomena (contextual boundness and communicative dynamism), it is possible to distinguish two parts of the sentence – topic and focus. These terms no longer concern the individual sentence elements as contextual boundness and communicative dynamism but are related to the larger parts of sentences.

Generally speaking, between topic and focus, there is a *relation of aboutness* – focus says something about the topic (cf. Hajičová, Partee and Sgall, 1998).

A simple example of topic and focus can be demonstrated as follows (topic is in plain text, focus is printed in bold):

(91) *He*$_{\text{t}}$ ***looked***$_{\mathbf{f}}$ ***at MAGDA***$_{\mathbf{f}}$. (Škvorecký, 1986)

The sentence is about *him*, hence this is the sentence topic. The other part of the sentence (*looked at MAGDA*) is a statement *about him*, i.e. sentence focus.

In a more detailed description, we can characterize topic as the part of a sentence that consists of all contextually bound sentence items directly dependent on their main governing verb. These items can also be further modified by other sentence members (e.g. by attributes) that can be contextually bound or non-bound – all such modifiers are also a part of topic.

At the same time, focus consists of all contextually non-bound sentence items directly dependent on their main governing verb. Also these items can be further modified by other sentence elements (like by attributes) that can be contextually non-bound or bound – all such modifiers are also a part of focus, see Example 92:

(92) (*I have two cats.*) *The black*$_{\text{c}}$ *one*$_{\text{t}}$ ***is***$_{\mathbf{f}}$ ***my***$_{\mathbf{t}}$ ***FAVORITE***$_{\mathbf{f}}$.

Example 92 demonstrates that the element *my* is a part of focus, though it is contextually bound.

Also the governing verb itself can be contextually bound or non-bound. If it is contextually bound, it is a part of topic; if it is contextually non-bound, it is a part of focus, see Example 93. For more details about the algorithm for detection of topic and focus, see, in particular, Sgall, Hajičová and Panevová (1986) and Zikánová, Týnovský and Havelka (2007).

(93) *He* ***looked***$_{\mathbf{f}}$ ***at MAGDA*** *while Magda looked*$_{\text{t}}$ ***at someone ELSE***.

The first occurrence of the governing verb *to look* is contextually non-bound and it is a part of focus. On the contrary, its other occurrence is contextually bound (deducible from the context) and therefore it is a part of topic.

In terms of communicative dynamism, topic is (as a whole) less dynamic than focus. At the same time, the individual items of topic have different degrees of communicative dynamism. The least dynamic item (i.e. the item with the lowest relative degree of importance) is called *topic proper*. Also the individual parts of sentence focus carry different degrees of communicative dynamism and the most dynamic item is called *focus proper* (in the spoken variant of the sentence, focus proper also carries the intonation centre), see Example 94:

(94) *Then*$_\text{t}$ *he*$_\text{t}$ ***RECOGNIZED***$_\textbf{f}$ *them*$_\text{t}$. (Škvorecký, 1986)

In Example 94, topic proper is the sentence element *he* (Actor is very often the item that is spoken about) and focus proper is the predicate because it carries the most important information.

In the next example, we present the sentences from previously used Example 88 once more – this time not only with values of contextual boundness of individual sentence items but also with marking of topic and focus in each sentence. Topics are written as plain text and focuses are printed in bold.

(95) *Across the river*$_\text{c}$ *Magda*$_\text{t}$ *and Kovarik*$_\text{t}$ *could now*$_\text{t}$ ***see***$_\textbf{f}$ ***a FIRE***$_\textbf{f}$ ***with two***$_\textbf{f}$ ***figures***$_\textbf{f}$ ***beside it***$_\textbf{t}$. *When they*$_\text{t}$ *moved*$_\text{t}$ *closer*$_\text{f}$, *they*$_\text{t}$ ***could make***$_\textbf{f}$ ***out two***$_\textbf{f}$ ***white***$_\textbf{f}$ ***HORSES***$_\textbf{f}$ ***against the background***$_\textbf{f}$ ***of the dark***$_\textbf{f}$ ***bushes***$_\textbf{f}$. *Then*$_\text{t}$ *he*$_\text{t}$ [*Kovarik*] ***RECOGNIZED***$_\textbf{f}$ *them*$_\text{t}$. (Škvorecký, 1986)

All sentences contain focus but not all of them also have topic (there are e.g. some sentences that are formed only by focus proper). The *focus proper* is an obligatory part of every sentence. It brings the most important information – the main message. Without the main message, it would not make sense to use the sentence in authentic communication. On the other hand, the *topicless sentences* (sometimes called hot news) are not rare. Such sentences are typically headlines or first sentences of the text presenting some new objects on the scene or very short sentences, see Examples 96–99.

(96) *How*$_\text{f}$ *Colorado*$_\text{f}$ *State*$_\text{f}$ *Won*$_\text{f}$ *by Losing*$_\text{f}$ *Jim*$_\text{f}$ *McElwain*$_\text{f}$ *to FLORIDA*$_\text{f}$.

(97) *Once upon a time*$_\text{f}$ *there were*$_\text{f}$ *three*$_\text{f}$ *FROGS*$_\text{f}$.

(98) *ATTENTION*$_\text{f}$*!*

(99) *Page*$_\text{f}$ *45*$_\text{f}$.

5.5 Detection of Topic and Focus

As we have seen in all previously mentioned examples, the main issue in recognizing topic and focus in sentences is an appropriate identification of the contextual boundness of individual sentence items. At the same time, topic and focus can also be detected by using operational criteria, the following two being most useful: the so-called *question test* and *test by negation*.

5.5.1 Question test

The range of the focus can be reliably detected by the *question test*. Its formulation assumes that for every sentence it is possible to determine a set of questions which can be appropriately answered by the given sentence (with its given surface word order and given realization of intonation), see Example 100.

(100) *Tomorrow, I will read a MAGAZINE.*

For the sentence realization with the intonation centre placed at the item *magazine*, examples of appropriate questions are *What will you do tomorrow?* or *What will you read tomorrow?* On the contrary, an example of an inappropriate question is *When will you read a magazine?*

Each appropriate question must fully represent the relevant features of the context in which the sentence may be used. However, it should be noted that it is an artificial pair of question and answer and that it is not a natural dialogue.

The aim of the test is to identify to which part of the sentence (topic or focus) the given elements belong. In the test only the appropriate questions are used. Those sentence elements that are contained in each of the appropriate questions belong to the sentence topic; those of its elements that are not found in any given set of appropriate questions belong to its focus; those elements of the sentence which only occur in some of the appropriate questions (but not in all of them) create the potential range of the sentence topic or focus.

The application of the question test is illustrated in Example 101. We also formulated a set of possible questions – for each question, we indicated which elements from the response to the question are not included in the question itself.

(101) *Kids are playing with SNOW.*

(101a) *What are the kids playing with?* (*… With SNOW.*)

(101b) *What are the kids doing?* (*… They are playing with SNOW.*)

The member *snow* is not present in any of the created sentences – this member is thus determined as the sentence focus proper. On the other hand, the item *kids* occurs in both of them – this item is therefore the topic proper. Other sentence elements stand between the two terminal points (on a scale of communicative dynamism) and they are the potential range of the sentence topic or focus depending on the appropriate questions.

It is also possible to imagine a context indicated in question below where none of the sentence elements are included in the question. If we accept this possibility, Example 101 would be understood as a sentence without topic, i.e. as hot news.

(101c) *What is going on?* (*... Kids are playing with SNOW.*)

While assembling the set of possible questions, we may see that according to the context (represented in questions), the tested sentence can have three possible meanings, i.e. three possible interpretations[40] of topic–focus articulation (focus is printed in bold in every example).

(102a) *What are the kids playing with?*
*Kids are playing **with SNOW**.*

(102b) *What are the kids doing?*
*Kids **are playing with SNOW**.*

(102c) *What is going on?*
***Kids are playing with SNOW**.*

5.5.2 Test with negation

Besides the question test, we can also use tests associated with negation as an operational criterion for determining certain aspects of topic–focus articulation. Hajičová (1973) consistently deals with this relationship of topic–focus articulation and negation (see e.g. also Hajičová, 1975). She concludes that in the primary case, the scope of negation is consistent with the focus of the sentence; the relation of the focus to the topic is thus negated (the sentence says that the focus is not true in relation to the topic).

The *test with negation* can be complemented by the notion of *possible continuations* as introduced by Chomsky (1969). His approach is based on the fact that in a natural continuation of the sentence, focus may contain parts of sentences that may be replaced by other parts, standing in a similar position (e.g. after the conjunctions *but, rather*). Chomsky particularly exemplifies this idea for questions and negative sentences but it is possible to also use it for the affirmative or negative form of Example 101 with its natural continuations, see Example 103.

(103) *Kids are (not) playing with SNOW.*

(103a) *Kids are not playing with SNOW but **with sand**.*

(103b) *Kids are not playing with SNOW but (they) **are watching TV at home**.*

(103c) *Kids are not playing with SNOW but **parents are buying sweets**.*

The results obtained by Chomsky's method are the same as those from the question test. According to the context, there are also three possible interpretations of topic–focus articulation (with the given intonation).

[40] with the given intonation

5.6 Representation of TFA in the Prague Dependency Treebank

The phenomenon of topic–focus articulation (under different names) is captured in several annotated corpora, e.g. the Potsdam Commentary Corpus (Stede and Neumann, 2014); the ANNIS Database (Annotation of Information Structure; Dipper, Götze and Skopeteas, 2007); the Muli corpus (Baumann et al., 2004); the Switchboard Corpus (Calhoun et al., 2005); the DannPASS (Danish Phonetically Annotated Spontaneous Speech; Paggio, 2006) or the Penn TreeBank (Bohnet, Burga and Wanner, 2013).

Every corpus representation of topic–focus articulation is unique and often differs significantly from other ones. Both the technical approaches and the annotated features are different. In this section, we introduce practical issues connected with annotating topic–focus articulation in the Prague Dependency Treebank using the theory of the Functional Generative Description described above (see also Mírovský et al., 2013).

The information about topic–focus articulation in the PDT is based on the following two characteristics: (i) value of contextual boundness and (ii) value of communicative dynamism. The sentence division into topic and focus is not explicitly annotated but is well deducible from the two annotated phenomena.

The annotation of topic–focus articulation proceeds on the tectogrammatical sentence layer, i.e. on the layer of deep syntax, and it is done on the dependency trees (which is unique within the other corpus annotations of topic–focus articulation).

The tectogrammatical trees also contain reconstructed sentence items, i.e. items (nodes) that are deleted in the surface sentence structure (see Chapter 1) and therefore it is possible to also annotate nodes present only in the deep sentence structure (e.g. elided subjects) but clearly participating in topic–focus articulation – elided sentence participants are usually contextually bound. On the tectogrammatical layer, other elements elided in the surface are also captured and the topic–focus articulation is annotated by all of them – these are e.g. obligatory participants from the valency frame of verbs or actual ellipses like *red* [*wine*] *and white wine*.

5.6.1 Annotation of contextual boundness in the PDT

In the first step of annotation, we evaluate each node (relevant for topic–focus articulation)[41] of a tree in terms of contextual boundness. In this respect, we distinguish sentence elements that are (i) contrastive contextually bound (marked as *c* and highlighted in green in our figures) (ii) non-contrastive contextually bound (marked as *t* and highlighted in white) and (iii) contextually non-bound (marked as *f* and highlighted in yellow), see Example 104 and Figure 5.1.

[41] The *tfa* value is not assigned e.g. to the technical root of the sentence or to the nodes representing coordinating conjunctions.

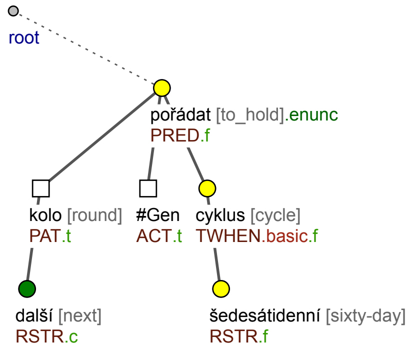


Figure 5.1: Example of annotation of topic–focus articulation in the PDT – contextual boundness

(104) (*Jak dále řekl Z. Škuta, trvalo první kolo 90 dnů.*) *Další*$_{\text{c}}$ *budou pořádána*$_{\text{f}}$ *v šedesátidenních*$_{\text{f}}$ *CYKLECH*$_{\text{f}}$*.* (PDT)

lit. (*As Z. Škuta further said, the first round took 90 days.*) *Next*$_{\text{c}}$ *will_be held*$_{\text{f}}$ *in sixty-day*$_{\text{f}}$ *CYCLES*$_{\text{f}}$*.*

(*As Z. Škuta further said, the first round took 90 days.*) *Next*$_{\text{c}}$ *rounds will be held*$_{\text{f}}$ *in sixty-day*$_{\text{f}}$ *CYCLES*$_{\text{f}}$*.*

Figure 5.1 captures the sentence from Example 104 in the dependency tree with the topic–focus articulation annotation (the previous context is indicated in brackets). In the tectogrammatical tree, we can find six TFA-relevant nodes – four are also expressed in the surface sentence structure (*next*, *to hold*, *sixty-day*, *cycle*) and two are present only in the deep sentence structure, i.e. on the surface, they are elided (*round* and so-called General Actor depending on the verb *to hold*).

Two nodes are non-contrastive contextually bound (*round* and General Actor) – the author considered them to be known and activated to such an extent that he or she had no need to express them at all. One node is contrastive contextually bound (*next*). *Next* expresses contrast to the previously mentioned *first* (*round*) and in a spoken variant of the sentence, it would carry the contrastive stress. Other three nodes (*to hold*, *cycle*, *sixty-day*) are contextually non-bound. They bring information non-deducible from the previous context. The sentence can also be interpreted as an answer to the question *What about the next rounds?* representing the known context.

The division of the sentence into topic and focus can be derived from the performed annotation of contextual boundness. As indicated above, topic consists of all contextually bound nodes directly dependent on the governing verb and by all nodes modifying these immediate dependents on the verb. Focus consists of all contextually non-bound nodes directly dependent on the governing verb and also of all nodes modifying these direct verb modifications.

In the dependency tree, there are three nodes directly dependent on the governing verb *to hold* (*round*, General Actor, *cycle*) and two further (lower) modifications (*next*, *sixty-day*). According to the definition, topic is General Actor (i.e. *somebody*) and *next rounds* and focus *will be held in sixty-day cycles*.

Example 105 and Figure 5.2 demonstrate that there are also authentic sentences that are as a whole contextually non-bound (the topicless sentences).

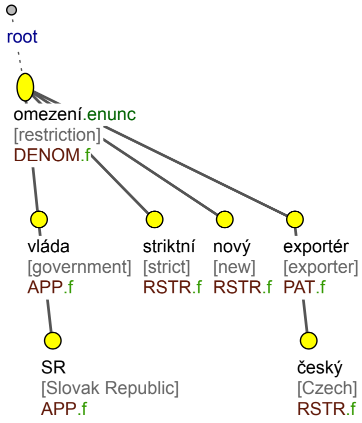


Figure 5.2: Topicless sentence from the PDT (Example 105)

(105) *Nová*$_\mathrm{f}$ *striktní*$_\mathrm{f}$ *omezení*$_\mathrm{f}$ *vlády*$_\mathrm{f}$ *SR*$_\mathrm{f}$ *proti českým*$_\mathrm{f}$ *EXPORTÉRŮM*$_\mathrm{f}$*.* (PDT)

New$_\mathrm{f}$ *Strict*$_\mathrm{f}$ *Government*$_\mathrm{f}$ *Restrictions*$_\mathrm{f}$ *of the Slovak*$_\mathrm{f}$ *Republic*$_\mathrm{f}$ *against Czech*$_\mathrm{f}$ *EXPORTERS*$_\mathrm{f}$*.*

Example 105, captured in Figure 5.2, is a heading of a newspaper article. All nodes carry some information non-deducible from the context; thus they are marked as contextually non-bound. This example illustrates that the whole sentence may have only the focus part and may lack the topic part.

5.6.2 Annotation of communicative dynamism in the PDT

The second step of annotation of topic–focus articulation in the PDT concerns communicative dynamism, i.e. ordering of nodes in the tree with respect to their communicative dynamism that grows from the left to the right. The deep ordering of the sentence elements in the PDT may be thus different from the surface word order of the given sentence.

The main rule for the communicative dynamism annotation in the deep word order is that in the dependency tree, the contextually bound nodes are placed to the left of the governing node, whereas the contextually non-bound nodes are placed to the right. The node that is placed in the rightmost position is the *focus proper* (i.e. the most dynamic part of the given sentence carrying the intonation centre).

The order of nodes in accordance with the communicative dynamism is observed only in the topic part of the sentence. In the focus, the nodes are ordered in accordance with their surface word order. The reason is that given the free word order (see Chapter 1) in Czech, the focus elements are supposed to also follow the scale of communicative dynamism in the surface sentence structure.

The annotation of communicative dynamism in the PDT is demonstrated in Example 106.

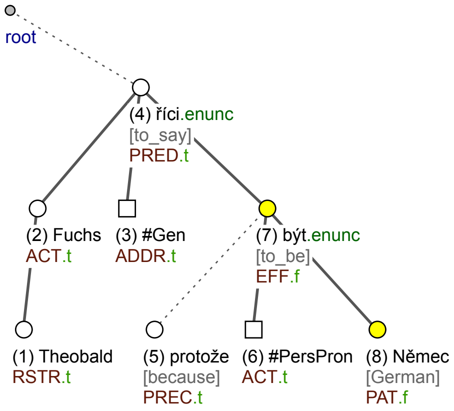


Figure 5.3: Annotation of topic–focus articulation in the PDT – communicative dynamism

(106) *„Protože jsme NĚMCI," řekl Theobald Fuchs.* (PDT)
"Because we are GERMANS," said Theobald Fuchs.

Figure 5.3 captures the sentence from Example 106 in the dependency tree of the PDT. According to the annotation of contextual boundness, the sentence topic consists of the part *Theobald Fuchs said* and focus *Because we are Germans*. In the surface deep order, focus precedes topic, which is against the principle of communicative dynamism. In Figure 5.3, the individual nodes are, therefore, shifted. Contextually bound nodes are shifted to the left from the governing verb *to say* and contextually non-bound nodes to the right. The most dynamic focus proper (*Germans*) is the last node in the tree.

In the figure, we may also see that communicative dynamism is also annotated on lower levels of nodes – e.g. the less dynamic node *Theobald* stands to the left of its more dynamic governing node *Fuchs*. The whole scale of communicative dynamism for Example 106 is: (1) *Theobald* – (2) *Fuchs* – (3) General Addressee (i.e. *to someone*) – (4) *said* – (5) *because* – (6) *we* – (7) *are* – (8) *Germans*.

5.7 Summary

In the present chapter, we have briefly introduced the theory of topic–focus articulation from the perspective of the Functional Generative Description. The fundamental features which the Functional Generative Description works with are contextual boundness and communicative dynamism. These two phenomena also serve as grounds for delimitation of topic and focus.

The theory of the Functional Generative Description also served as a basis for the annotation of topic–focus articulation in the Prague Dependency Treebank. Topic–focus articulation is annotated on the tectogrammatical (deep syntactic) layer of language in two steps – as contextual boundness and communicative dynamism in dependency trees. The division of sentences into topic and focus is not explicitly marked but it is clearly deducible from the annotation of contextual boundness and communicative dynamism.

The annotation of topic–focus articulation in the Prague Dependency Treebank belongs to the phenomena with very high inter-annotator agreement – despite the fact that the annotation of authentic texts depends to some extent on the annotator's interpretation (see above mentioned ambiguous sentences). The inter-annotator agreement in assigning the value to individual nodes in the annotation of topic–focus articulation was 82% (see Chapter 7).

There are still a few open questions. One of them is a further study of contrastive contextually bound nodes. During annotations of written texts, it turned out that the annotators are not sure in some cases whether the given node can or cannot bring the facultative contrastive stress. Yet, the possible occurrence of the contrastive stress is crucial in order to decide whether the sentence element is contrastive or non-contrastive contextually bound. In such cases, the annotators have to rely on their language consciousness and experience to some extent.

At the current stage, the PDT contains the most detailed annotation of topic–focus articulation (carried out on the largest language material) in Czech (and one of the largest in general) and thanks to annotations of other language phenomena (like discourse or coreference relations etc.), it is a comprehensive source for complex studies of text coherence as well as other language issues in interaction, see Chapter 13.

Data

6

Prague Dependency Treebank

Data to explore and exploit

Any theory (albeit possibly interesting) remains merely a thought experiment until it is subjected to the test of application in the real world. For linguistic theories, this means applying them to real language data. In our case, the analyses described in the theoretical chapters have been applied to the data of the Prague Dependency Treebank.

The Prague Dependency Treebank (PDT) is a corpus of continuous Czech texts mostly of the journalistic style, consisting of almost 50 thousand sentences annotated mostly manually at three layers of language description: *morphological*, *analytical* (surface syntactic structure), and *tectogrammatical* (deep syntactic structure). The Prague Dependency Treebank 3.0 (Bejček et al., 2013) is the latest version of the PDT, succeeding versions 1.0 (PDT 1.0; Hajič et al., 2001), 2.0 (PDT 2.0; Hajič et al., 2006), 2.5 (PDT 2.5; Bejček et al., 2011) and the Prague Discourse Treebank 1.0 (PDiT 1.0; Poláková et al., 2012b; Poláková et al., 2013).

The annotation scheme of the PDT has been derived from a solid, well-developed theory of a language description called Functional Generative Description (FGD; Sgall, 1967a; Sgall et al., 1969; Sgall, Hajičová and Panevová, 1986). The FGD framework (see also Chapter 1) was formulated as a generative description that was conceived of as a multi-level system proceeding from linguistic function (meaning) to linguistic form (expression), that is from the generation of the deep syntactico-semantic representation of the sentence through the surface syntactic, morphemic and phonemic levels down to the phonetic shape of the sentence.

The design of the annotation scenario of the PDT follows the above conception of the FGD in all the fundamental points, and most importantly:

- it is conceived of as a multilevel scenario including the underlying syntactico-semantic layer (tectogrammatical),
- the scheme includes a dependency based account of syntactic structure at both syntactic levels.

6.1 Layers of Annotation

In the PDT, the original text is represented at **the word layer** (w-layer),[42] where the text is segmented into documents and paragraphs and individual tokens are identified and associated with unique identifiers. At **the morphological layer** (m-layer), the sequence of tokens of the w-layer is divided into sentences. Several attributes are assigned to each token, the most important of which are the disambiguated *lemma* and morphological *tag*. A sentence at **the analytical layer** (a-layer) is represented as a dependency tree with labelled nodes and edges. Each token of the m-layer is represented at the a-layer by exactly one node of the tree and the dependency relation between two nodes of the a-layer is captured by an edge between them. The actual type of the relation is given as a function label of the edge, represented by the attribute *afun* at the dependent node with values such as *Pred* (*Predicate*), *Sb* (*Subject*), *Obj* (*Object*), or *Atr* (*Attribute*). Most of the edges represent dependency relations, while the rest stand for various linguistic or technical phenomena such as coordination, apposition, punctuation, etc.

At **the tectogrammatical layer** (t-layer), every sentence is also represented as a dependency tree with labelled nodes and edges. The tree reflects the underlying (deep) structure of the sentence. The nodes stand for auto-semantic words only (with some technical exceptions). Unlike the analytical layer, not all morphological tokens are represented at the tectogrammatical layer as nodes (for example, there are no prepositions there) and, on the other hand, some of the tectogrammatical nodes do not correspond to any morphological tokens (for example, the structure contains nodes representing elided subjects in pro-drop constructions). The edges of the tree represent relations between the nodes they connect; the type of the relation is indicated by the label of the particular edge, which is, similarly to the analytical layer, expressed at the dependent node, in the attribute *functor* with values such as *PRED* (*Predicate*), *ACT* (*Actor*), *ADDR* (*Addressee*), *MANN* (*Manner*), *LOC* (*Locative*), or *DIR* (*Direction*). *Grammatemes* are attached to some nodes; they provide information about the node that cannot be derived from the structure, *functor* and other attributes – for example number for nouns, modality and tense for verbs, etc. For every node representing a verb or a certain type of noun, a valency frame assigned to it can be detected by means of a reference to a valency dictionary. Within the context of annotation of the topic–focus articulation, (i) each node is assigned one of the three values of contextual boundness (attribute *tfa*): a node can be contextually bound, contrastive contextually bound, or contextually non-bound, and (ii) the (contrastive) contextually bound nodes of the tree are ordered in the left-to-right direction according to the assumed communicative dynamism. In total, there are 39 attributes assigned to every non-root node of the tectogrammatical tree. Based on the node type, only a certain subset of the attributes is necessarily filled in.

[42] This means that there are actually four layers in the PDT; however, only the higher three layers are called “annotation” layers.

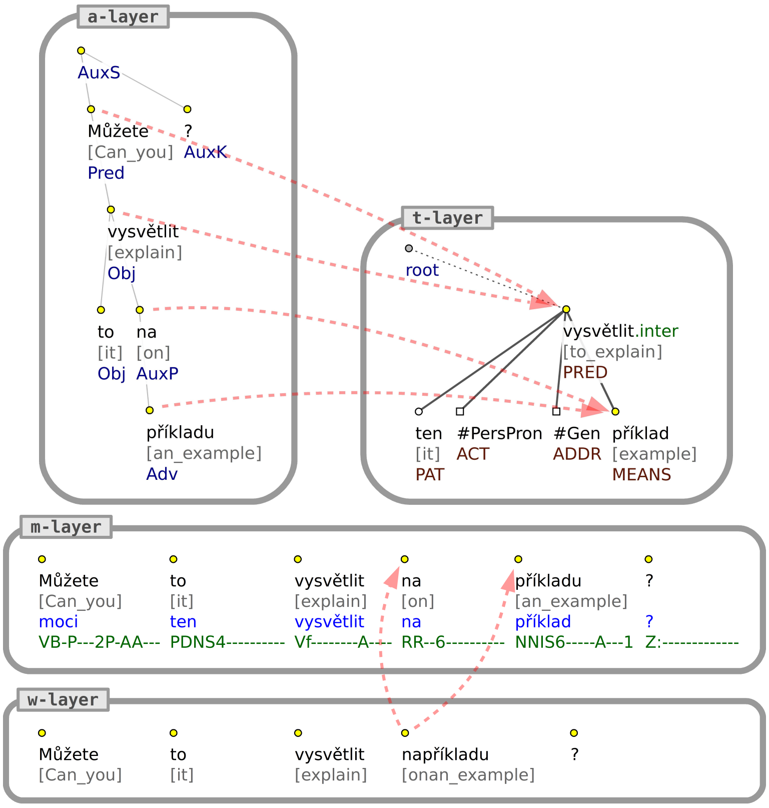


Figure 6.1: Interlinking of the PDT layers for the sentence from Example 107

(107) *Můžete to vysvětlit na příkladu?* (PDT)
lit. *Can_you it explain on an_example?*
Can you explain it with an example?

In Figure 6.1, layers of the PDT are illustrated on a simple Czech sentence from Example 107.[43] The arrows represent non-1:1 relations among tokens and/or nodes at the different layers; square nodes in the tree at the t-layer represent newly generated nodes – nodes without a surface counterpart.[44]

[43] It is a real example from the PDT, except for the misprint depicted in the figure at the word layer, which we have introduced (only) to the figure to demonstrate different roles of the w-layer and m-layer: Errors in the original text (such as the missing space in the figure between words *na* [*on*] and *příkladu* [*an example*]) are preserved at the w-layer and corrected at the m-layer.

[44] Especially at the t-layer, substantially more attributes are annotated at the t-nodes. For simplicity, only the tectogrammatical lemma (*t-lemma*) and *functor* are displayed in the figure.

In the figure, the words **například* [**onan_example*], here on purpose erroneously printed without a space, are represented as a single token at the w-layer. The error is corrected at the m-layer, where the words *na* [*on*] and *příkladu* [*an example*] are represented as two tokens with morphological lemmas and morphological tags assigned. At the a-layer, these two tokens are represented as two nodes, with the preposition *na* [*on*] governing the noun *příkladu* [*an example*], playing the role of adverbial (*afun* = *Adv*) to the verb *vysvětlit* [*explain*]. On the t-layer, these two nodes are represented by a single node with the tectogrammatical lemma *příklad* [*an example*] and *functor MEANS*.

6.2 Discourse Coherence Phenomena

Annotation of the topic–focus articulation (see Chapter 5), as well as annotation of the grammatical coreference and pronominal textual coreference (see Chapter 3) appeared already in 2006 in the PDT 2.0, as a part of the annotation of the tectogrammatical layer. Technically, the contextual boundness of the individual nodes is marked in the attribute *tfa*, with possible values *t* (contextually bound), *c* (contrastive contextually bound), and *f* (contextually non-bound). The left-to-right order of the (contrastive) contextually bound nodes in the tree, defined by numeric values of the attribute *deepord*, reflects the communicative dynamism of these sentence elements.[45] Coreferential relations are represented by a reference from one of the respective nodes (*start node*) to the other (*target node*), taking advantage of the fact that each node has a corpus-wide unique identifier.

The first version of annotation of the extended textual coreference (see Chapter 3), bridging anaphora (Chapter 4) and discourse relations (Chapter 2) was published in 2012 in the PDiT 1.0. Similarly to the pronominal textual coreference, they were annotated on top of the tectogrammatical layer, using a similar technical solution. The relation is represented as an arrow pointing from the start node to the target node, carrying additional information (most importantly the type of the relation; for discourse relations, the range of the arguments and a list of nodes that form the connective of the relation are also represented). In 2013 in the PDT 3.0, the annotation of textual coreference was further extended to also include coreference relations for pronouns of the 1st and 2nd person. Annotation of discourse relations was also updated and slightly extended (for details see Mírovský, Jínová and Poláková, 2014).

Figure 6.2 shows a graphical representation of annotation of discourse coherence phenomena in the two sentences from Example 108, as depicted in the tree editor TrEd (Pajas and Štěpánek, 2008), a primary tool used for the annotation of the PDT.[46]

[45] Contextually non-bound nodes are generally ordered in accordance with the surface order.

[46] The translations of the tectogrammatical lemmas are not a part of the PDT.

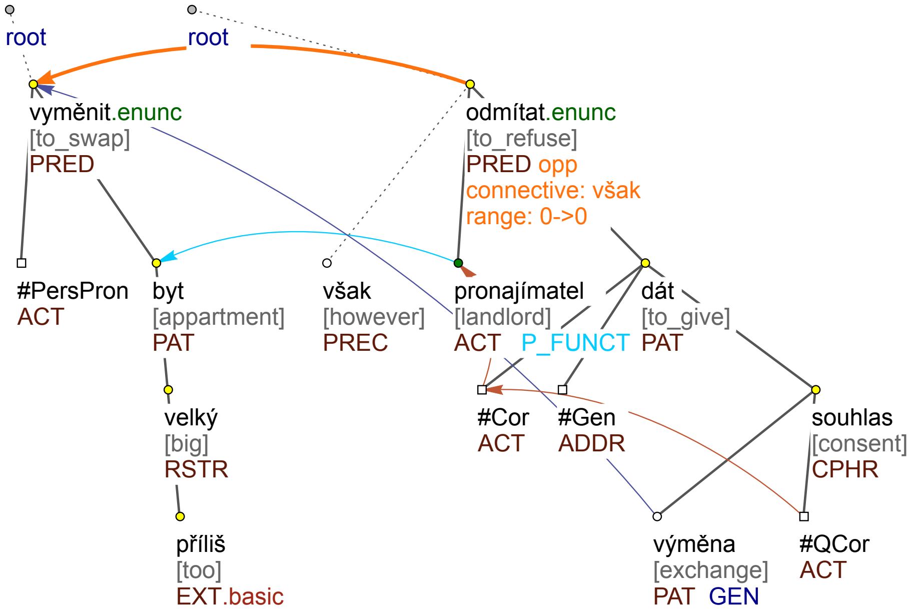


Figure 6.2: Discourse coherence phenomena annotated in the sentences from Example 108

(108) *Chtěl bych vyměnit příliš velký byt.*
Pronajímatel však odmítá dát k výměně souhlas. (PDT)

lit. *I would like to swap a too big appartment.*
The landlord however refuses to give for the exchange the consent.

I would like to swap an apartment that is too big.
But the landlord refuses to give his consent for the exchange.

In the figure, the thick orange arrow starting at the node *odmítat* [*to refuse*] and ending at the node *vyměnit* [*to swap*] denotes a discourse relation between two arguments represented by the nodes, i.e. in this case between the two subtrees of the nodes, as indicated by the range values (*0->0*)[47] at the start node. There is also information about the semantic type of the relation (*opp*, meaning *opposition*), and the surface representation of the connective (*však* [*however* or *but*]). In a similar way, dark blue arrows

[47] Range *0* means that the subtree of the given node represents the argument; any other number means that the argument consists of the subtree plus the given number of the subsequent trees (or the preceding trees if a negative number is used); other values for the range of an argument (*group*, *forward*, *backward*) express even more complex cases.

mark the textual coreference (in the figure between the nodes *výměna* [*exchange*] and *vyměnit* [*to swap*] with the type *GEN*, indicating a non-specific reference between the noun and the clause), and light blue arrows denote the bridging anaphora (in the figure there is an anaphoric relation of the type *P_FUNCT* (relation *individual–function*) between the nodes *pronajímatel* [*landlord*] and *byt* [*appartment*]). The remaining (dark red) arrows denote the grammatical coreference. The contextual boundness of the nodes is represented by their colour: contextually bound nodes are displayed as white circles (or squares), contrastive contextually bound nodes are depicted in green, and contextually non-bound nodes in yellow.

For the sake of completeness of our description of the PDT, let us mention also the annotation of multiword expressions that was published in the PDT 2.5. A list of multiword expressions appearing in the given sentence is kept at the technical root of the tectogrammatical tree in the attribute *mwes* and each multiword expression is assigned a type, such as *person*, *institution*, *location*, or *time* (Bejček and Straňák, 2010). Another type of annotation worth mentioning is the annotation of genres that was published as a part of the PDT 3.0. Each document is assigned a single value of the attribute *genre*,[48] such as *news*, *review*, *topic interview*, or *letter*, see Chapter 2, Table 2.2 for a full list (Poláková, Jínová and Mírovský, 2014).

The total number of documents in the PDT annotated at all three annotation layers is 3,165, amounting to 49,431 sentences and 833,193 nodes.[49] All the discourse coherence phenomena have been annotated on the same data, i.e. on top of the whole tectogrammatical layer. For the purposes of natural language processing (NLP) applications (such as machine learning), the data are divided into ten parts. Eight of them (*train-1* – *train-8*) are designated as training data, *dtest* is designated as development test data, and *etest* is meant to serve as evaluation test data. Honoring this division and designation of the data in the subsequent case studies presented in this book, some measurements were performed on the training data only (8/10 of the whole data) or on the training data and development test data (9/10 of the whole data), leaving the whole test data or at least the evaluation test data "unobserved." Generally, only overall statistics that are not supposed to be used in automatic NLP methods are presented for the whole PDT data. In each case study, it is always specified on which part of the PDT data the measurements were performed.

[48] kept in the t-layer file

[49] These are the numbers for documents annotated at all three layers. There are additional documents in the PDT annotated only up to the analytical layer and even more documents annotated only at the morphological layer.

7

Inter-Annotator Agreement

Since the beginning of the annotation of the Prague Dependency Treebank (PDT) in the late 1990's, the inter-annotator agreement has been measured for individual annotation tasks. Studying the disagreements helped detect errors in the annotations, improve the annotation guidelines and find difficult phenomena from the annotation perspective. In this chapter, we have put together information on these measurements from various published papers and sorted them by the layer of annotation or level of language abstraction, proceeding from the morphological layer, to higher layers of single sentence annotation, to phenomena crossing the sentence boundary, and finally to document-level phenomena, namely annotation of genres of documents. In some places, we also offer numbers of inter-annotator agreement measurements from other similar projects and other languages.

Naturally, measurements of the inter-annotator agreement can only be done on data annotated independently by two (or more) annotators in parallel. As annotations are cost-demanding tasks, usually most of the data are annotated by one annotator only and just a small part is annotated in parallel by two annotators, in order to measure and study the inter-annotator agreement. These measurements need to be made before or shortly after the real annotations start, as discrepancies between the annotators can reveal flaws in annotation instructions or misunderstanding of the instructions, which both need to be resolved as quickly as possible. Subsequent measurements of the inter-annotator agreement should be performed regularly during the whole annotation process, without the annotators knowing when the check will be carried out or which part of the data will be used for the measurements. This ensures that the annotators continue to work with care and high thoroughness; and as the annotators become more and more experienced, an improvement in the quality of the annotations may be detected.

For different types of annotation tasks, different measures of inter-annotator agreement need to be used. For classification tasks, where positions in the data to be annotated are given, such as morphologically ambiguous words in the task of disambiguation of morphological analysis, a simple ratio of positions where the annotators agreed on the assigned value is measured, and is usually given in percents. For more complex tasks, where the selection of the places to be annotated is also a part of the annotators' decision, such as annotation of bridging anaphora or discourse relations, we use the F1-measure, which is a harmonic mean between precision and recall;[50] F1 can be given in percents or (equivalently) as a number between 0 and 1. Assigning a type of such a relation for places where the annotators agreed on its existence is, again, a simple classification task for which we give the simple ratio of agreement.

It is very difficult to say what is a good inter-annotator agreement. It very much depends on the complexity of the annotations (which we try to demonstrate in this chapter) and also on the type of data. Part of the inter-annotator agreement can also be random, especially, for example, for classification tasks where one of the assigned values occurs much more often than the other values. Cohen's κ (kappa; Cohen, 1960) tries to subtract this random agreement from the measurement, i.e. it captures the level of agreement that is beyond chance; it can be given in percents or (equivalently) as a number between 0 and 1.

7.1 Within a Single Sentence

On **the morphological layer in the PDT**, the annotation task was to choose the correct lemma and morphological tag for each ambiguous token in the input text, i.e. for each token where the morphological analysis had offered several options. The disambiguation of the morphological analysis was done in parallel by pairs of annotators on (atypically) the whole PDT data. The inter-annotator agreement for the assignment of the correct morphological tag to words with an ambiguous morphological analysis was 95% (Bémová et al., 1999); if the unambiguous words are also counted, the agreement is 97% (Hajič, 2006). Discrepancies between the two annotators were later resolved by an arbiter – a third annotator.

We can compare our project to the inter-annotator agreement measurement during the annotation of the German corpus NEGRA, as reported by Brants (2000). Their agreement in the part-of-speech annotation was 98.57%. Note that the size of their part-of-speech tagset was 54 tags, while for Czech, there are almost 5 thousand different morphological tags.

The analytical layer in the PDT captures the surface syntax of the sentence. The annotation task was to establish dependencies between the words, i.e. to choose a governor for each word, and to assign a type of the dependency – the analytical function (attribute *afun*), such as *Pred* (*Predicate*), *Sb* (*Subject*), *Obj* (*Object*), or *Atr* (*Attribute*). For this type of annotation, two types of inter-annotator agreement are usually measured – *the unlabelled attachment score* reflects the agreement in choosing the governor for each node, and *the labelled attachment score* reflects the agreement both

[50] In annotation evaluation tasks, *precision* (P) is counted as a ratio between cases correctly marked by an annotator and all cases marked by the annotator; *recall* (R) gives the ratio between cases correctly marked by an annotator and all cases to be marked. In measuring precision and recall of the inter-annotator agreement, the annotation of one of the annotators is tested against the annotation of the other one (which is for the moment considered "correct"); if the roles of the annotators are swapped, the values of precision and recall also swap. F1-measure, which is counted as their harmonic mean ($F1 = 2 * P * R/(P + R)$), is therefore a symmetric measure of the agreement between the annotators.

in the choice of the governor and the assignment of the type of dependency. Alternatively, instead of the labelled attachment score, a ratio of agreement in assigning the type of dependency for dependencies the annotators agreed on may be given. As far as we know, no measurements of the inter-annotator agreement have been published for the analytical layer in the PDT. The NEGRA corpus is again an example of a similar project abroad. Brants (2000) reports the inter-annotator agreement (F-measure) for the unlabelled structural annotation as 92.43%, and for the labelled structural annotation (labelled nodes with 25 phrase types and labelled edges with 45 grammatical functions) as 88.53%. (For comparison, there are 28 analytical functions in the PDT.)

The tectogrammatical layer in the PDT captures the deep syntax of the sentence. The inter-annotator agreement measurements were performed during the annotation of the PDT (most of the numbers from this layer, unless specified otherwise, come from Hajičová, Pajas and Veselá, 2002) and of the Czech part of the Prague Czech-English Dependency Treebank (PCEDT; Mikulová and Štěpánek, 2010), which is based on the same annotation scenario. The annotation task was much more complex than for the analytical layer and consisted of several subtasks; the most important element was again to establish dependencies between nodes and to assign a type for each dependency – the attribute *functor* with values such as *PRED* (*Predicate*), *ACT* (*Actor*), *ADDR* (*Addressee*), *MANN* (*Manner*), *LOC* (*Locative*), or *DIR* (*Direction*). In total, there are 67 possible *functors* in the PDT. The agreement in establishing the correct dependency between pairs of nodes (i.e. the establishment of dependency links together with the determination which member of the pair is the governor) was 91% in the PDT, and 88% in the PCEDT. The agreement in assigning the correct type to the dependency relation (the tectogrammatical functor) was 84% in the PDT, and 85.5% in the PCEDT.

From other annotation subtasks at the tectogrammatical layer, we should mention the agreement in assigning sentence modality for 268 complex cases of coordinated clauses in the PDT (annotated in the PDT 3.0), which was 93.7% with Cohen's κ 0.89 (Ševčíková and Mírovský, 2012). The agreement in assigning a value to individual nodes in the annotation of the topic–focus articulation, i.e. the assignment of the values *contextually bound*, *contrastive contextually bound* or *contextually non-bound* within the attribute *tfa* (discussed in Chapter 5), was 82% (Veselá, Havelka and Hajičová, 2004). Agreement in linking the tectogrammatical nodes to their counterparts from the analytical layer in the PCEDT was 96% for the lexical counterparts and 93.5% for the auxiliary nodes.

In the task of marking multiword expressions in the data (which was done on top of the tectogrammatical layer for the PDT 2.5), the authors used their own version of weighted Cohen's κ and report the agreement above chance of 0.644 (Bejček and Straňák, 2010).

7.2 Crossing the Sentence Boundary

Three phenomena that cross the sentence boundary have been annotated on top of the tectogrammatical layer of the PDT data, namely the textual coreference (discussed in Chapter 3), the bridging anaphora (Chapter 4) and the discourse relations (Chapter 2). For the textual coreference and the bridging anaphora, the annotation task consisted of detecting arguments of the relation and assigning its type; the argument is always represented by a single tectogrammatical node and consists of the node and its subtree. For the discourse relations, the task was to find a connective of a relation, detect the arguments and assign a type of the relation; the argument here is also represented by a tectogrammatical node (standing for the whole subtree) but can be more complex – it can consist of several trees (sentences) or a part of a tree that does not have a form of a single subtree.

For all these three phenomena, ten subsequent measurements of the inter-annotator agreement were performed during the whole annotation process. For the textual coreference and the bridging anaphora, altogether 1,606 sentences (39 documents) were annotated in parallel by two annotators, and for the explicit inter-sentential discourse relations, the agreement was measured on 2,084 sentences (44 documents).

To evaluate the inter-annotator agreement in these annotations, several measures were used. The connective-based F1-measure (Mírovský, Mladová and Zikánová, 2010) was used to measure the agreement in the recognition of a discourse relation, the chain-based F1-measure was used for measuring the agreement on the recognition of a coreference or a bridging relation. A simple ratio and Cohen's κ were used for measuring the agreement in the relation types in cases where the annotators recognized the same relation (see also Poláková et al., 2013).

In the connective-based F1-measure, we consider the annotators to be in agreement in recognizing a discourse relation if the two connectives they mark (each of the connectives marked by one of the annotators) are at least partially overlapping (technically, a connective is a set of tree nodes, i.e. the condition is for the two sets of nodes to have a non-empty intersection).[51] For details, see Jínová, Mírovský and Poláková (2012a). In the chain-based F1-measure, we consider the annotators to be in agreement in recognizing a coreference or a bridging relation if two nodes connected by an arrow by one of the annotators have also been connected by the other annotator; coreference chains are taken into account, i.e. it is sufficient for the agreement if the arrow starts (or ends) in a node that is coreferentially connected (possibly transitively) with the node used for the relation by the other annotator.

The graph in Figure 7.1 shows results of ten subsequent measurements of the inter-annotator agreement performed during the two years of annotation of inter-sentential discourse relations in the PDT (taken from Jínová, Mírovský and Poláková, 2012a).

[51] For example, one of the annotators marks the expression *a proto* [*and therefore*] as a connective, while the other annotator marks only the word *proto* [*therefore*]; these two expressions are overlapping, thus we consider the annotators to agree on the presence of a discourse relation.

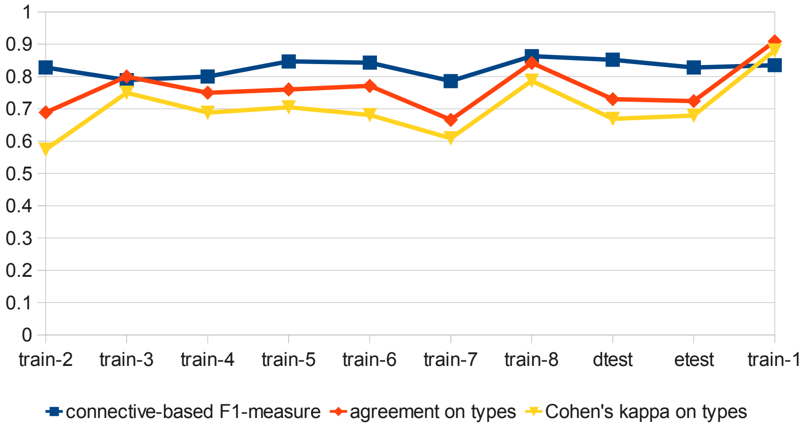


Figure 7.1: Subsequent measurements of the inter-annotator agreement for inter-sentential discourse relations

Each measurement was carried out on approx. 200 sentences (3 to 5 documents). The curves indicate generally consistent results, with possible slight improvements. Note that the first part of the PDT, *train-1*, appears at the far right end of the graph – since the annotators were not well enough trained when it was annotated for the first time and the annotation instructions were not entirely completed, this part of the data was re-annotated at the end of the annotations.

In a small probe of annotating implicit inter-sentential discourse relations (performed on 96 sentences from 3 documents from the PDT), the task proved to be highly challenging – the annotator's agreement in selecting a type of implicit discourse relations between adjacent sentences was less than 60%.

Table 7.1 shows overall results of the inter-annotator agreement measurements for all three annotated phenomena, performed on all data annotated in parallel by two annotators. Comparison of these numbers with other similar projects is difficult, as the projects usually use different annotation schemes and different scores. Nevertheless, to get a general idea, we can look at some numbers from similar projects. The simple ratio agreement for types in discourse relations (0.77 on all parallel data, see the third column of Table 7.1) is the closest measure to the way of measuring the inter-annotator agreement used on subsenses in the Penn Discourse Treebank 2.0, as reported in Prasad et al. (2008). Their agreement was 0.8. In the annotation of coreference relations in OntoNotes, the inter-annotator agreement on English was 0.81 for newspaper texts and 0.78 for magazine texts. For Chinese, the agreement achieved was 0.74 for newspaper texts and 0.75 for magazine texts (reported in Pradhan et al., 2012). These numbers can be compared with our chain-based F1 measure (0.72 in the second column of Table 7.1), as it is similar to the MUC-6 score they used. As for the

Relation	F1	Agreement in types	Cohen's κ
discourse	0.83	77%	0.71
textual coreference	0.72	90%	0.73
bridging anaphora	0.46	92%	0.89

Table 7.1: Overall inter-annotator agreement for discourse, coreference and bridging anaphora

bridging anaphora, we can compare our chain-based F1-measure (0.46 in the second column of Table 7.1) to F1-measure in recognition of bridging relations reported for the annotation of the COREA corpus (Dutch texts); their agreement on newspaper texts was 0.39 (Hendrickx, De Clercq and Hoste, 2011).

7.3 At the Document Level

The only annotation in the PDT performed at the document level was the annotation of genres, published in the PDT 3.0. The annotation task consisted of reading the document and assigning to it one of 20 possible values – genres such as *essay*, *news*, *comment*, *personal interview*, or *letter* (see Chapter 2, Table 2.2 for a full list). The inter-annotator agreement (reported in Poláková, Jínová and Mírovský, 2014) was measured between pairs of annotators and varied (depending on the given pair of annotators) around 60% (simple agreement ratio), with Cohen's κ around 52%. Table 7.2 shows a part of distributions of genres in different parts of the data annotated by four annotators (altogether there were 8 annotators), indicating that the annotator A3 probably misunderstood the annotation instructions, namely the definition of genres *description* and *news*, as their ratio in the annotated data is opposite to the ratio of these two genres in the annotations of the other annotators. These two genres in the data of the annotator A3 were subsequently re-annotated by another annotator to fix the problem, which was only revealed thanks to the analysis of the individual annotators' annotations.

7.4 Summary

Measuring the inter-annotator agreement and studying discrepancies between annotators repeatedly proved to be an indispensable part of the annotation process of the PDT. Not only is it necessary for ensuring a high quality annotation (for reasons mentioned above) but it may even reveal shortcomings in the underlying linguistic theory. It is the only way to establish and enumerate the difficulty of a given annotation task and to set a higher boundary for the accuracy we can expect from automatic methods of annotation.

A2		A3		A4		A5	
news	147	**description**	118	*news*	179	*news*	157
comment	50	comment	59	sport	43	sport	40
sport	36	sport	41	**description**	26	caption	29
description	22	essay	35	review	22	**description**	28
caption	20	*news*	28	comment	17	comment	23

Table 7.2: Comparison of (parts of) distributions of genres annotated by four annotators

Table 7.3 summarizes most of the inter-annotator agreement measurements from this chapter. Even if the numbers for the different tasks often cannot be directly compared – as they measure principally different phenomena, use different methods of evaluation and sometimes annotate different (types of) data in a different way – we can still use them to observe several tendencies (see also Mírovský and Hajičová, 2014).

The most obvious tendency is the increasing difficulty to achieve high values of the inter-annotator agreement, as we go deeper in the abstraction of the language description, from the morphological layer to surface syntax layer, to deep syntax layer and to inter-sentential relations. It is also quite clear from the table that recognizing presence of a textual coreference relation is easier than that of a bridging relation. For both textual coreference and bridging anaphora, it is more difficult to find the relation than to select its type – once the presence of the relation is agreed upon, the annotators are able to assign its type with high accuracy. For discourse relations, however, assigning the type of a relation seems to be more difficult than recognizing its presence.

It seems to be clear from the table that we cannot find a simple relation between the number of possible values for a classification annotation task and the accuracy of the annotation. For morphology, the numbers for Czech and German might support the idea that a smaller tagset leads to a higher agreement. The same holds for comparison of discourse relations with anaphoric relations, but not for textual coreference vs. bridging anaphora. From this perspective, we would also expect higher agreement for the annotation of topic–focus articulation (with only three possible values). The assumption might be true for various sizes of possible values for similar tasks (the same layer of annotation, the same level of language abstraction), but for different layers of annotation, other factors like subjectivity or vagueness of the phenomena in question seem to play a more important role.

Annotation task	Agreement (%)
morphology (Czech, 5 thousand tags)	97
morphology (German, 54 tags)	98.6
surface syntax (German, unlabelled structural annotation)	92.4
surface syntax (German, labelled structural annotation, 25 phrase types and 45 grammatical functions)	88.5
deep syntax (Czech, unlabelled structural annotation)	91
deep syntax (Czech, assigning the type of dependency, 67 functors)	84
topic–focus articulation (Czech, assigning contextual boundness, 3 values)	82
discourse relations (Czech, recognizing a presence of an (explicit) inter-sentential discourse relation)	83
discourse relations (Czech, assigning one of 23 types to explicit relations)	77
discourse relations (Czech, assigning one of 23 types to implicit relations)	60
textual coreference (Czech, recognizing a presence)	72
textual coreference (Czech, assigning one of 2 types)	90
bridging anaphora (Czech, recognizing a presence)	46
bridging anaphora (Czech, assigning one of 9 types)	92
genres of documents (Czech, 20 genres)	77

Table 7.3: Overview of a selected number of inter-annotator agreement measurements at different annotation layers. Please note that the numbers represent different measures and cannot be simply compared.

8

Searching in the PDT

Any large and richly annotated treebank would be of limited use if there was no way to effectively mine information from it, i.e. search for various phenomena that occur in the language and that have been annotated in the data. And if it is to be of value not only to computer scientists but also to (both theoretical as well as empirical) linguists, the search process needs to be simple and intuitive. The Prague Dependency Treebank is a very good example of a richly annotated treebank that poses a challenge for search tools. It contains annotations of several layers with non-trivial relations between some of them and with links to external resources (lexicons). For a manually annotated treebank, it is fairly large (50 thousand sentences annotated at all layers). The annotation is highly complex (the annotation guidelines for the tectogrammatical layer alone consist of more than twelve hundred pages). A tool that would allow for searching in and studying all annotated phenomena in the PDT has to be powerful in terms of the query language but simple to understand and use. Mírovský (2009) offers a study of what features a query language has to possess in order to be powerful enough for the PDT.

The PML-Tree Query (PML-TQ; Pajas and Štěpánek, 2009) is an advanced client–server system for searching in the Prague Dependency Treebank and other linguistically annotated treebanks encoded in the Prague Markup Language (PML; Hana and Štěpánek, 2012).[52] It offers a powerful query language with an intuitive, graphically oriented way of query creation.

Queries in the PML-TQ can be created both in a textual form and graphically. The basic (and simplified) idea of the system is such that a user draws a tree that should be included in a result tree as its subtree. The system processes the query and displays result trees one by one (if there are any), along with the context. The query language allows to define properties of tree nodes and relations among them (relations such as dependency, transitive dependency, left-right order, etc.) inside or between sentences and also across layers of annotation. Information from dictionaries (such as valency lexicons) can be easily incorporated. Negation and arbitrary logical constraints can be used in the queries. Results of the corpus search can be viewed one by one along

[52] Many existing treebanks have been transformed to the PML format and are searchable in the PML-TQ, including, for example, the Penn Treebank or the TIGER corpus. Also, in the project HamleDT (Zeman et al., 2014, see also https://ufal.mff.cuni.cz/hamledt), currently 30 dependency treebanks (or dependency conversions of other treebanks) have been harmonized into the same annotation style and are also searchable in the PML-TQ. For the list of treebanks available in the PML-TQ, see http://lindat.mff.cuni.cz/services/pmltq/.

with the context for a thorough inspection, or further processed with so-called output filters to produce statistical overviews. A detailed documentation can be found on the internet,[53] here we offer a simple introduction to the principal parts of the PML-TQ query language (Section 8.1) and show its usage on a set of illustrative discourse-related examples (Section 8.2). Section 8.3 gives technical details on how to download the PDT or how to access the public search server for the PDT.

8.1 Basics of the PML-TQ Language

8.1.1 Node selection

Values of attributes of a node can be set using several operators, mainly '=' for the equality relation, '~' for a regular expression, and 'in' for selection from a set of values. Each of these operators can be negated using the prefix '!'.

In Example 109, we are looking for sentences with an *Actor* atypically not expressed by a noun, i.e. for sentences like *It is difficult to live alone*, where the *Actor* (*to live*) is expressed by the infinitive verb form. The query consists of one tectogrammatical node and defines three of its attributes: The *functor* has to be an *Actor* (*ACT*), the semantic part of speech must not be a noun (i.e. it does not start with *n*), and the node does not have a substitute *t_lemma* (i.e. a *t_lemma* starting with #).[54]

(109a) The textual form of the query:

```
t-node
  [ functor = "ACT", gram/sempos !~ "^n", t_lemma !~ "^#" ];
```

(109b) The graphical form of the query:

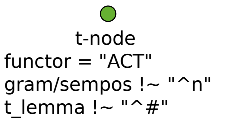


Figure 8.1 shows the tectogrammatical representation of one of the possible results. The matching node in the tree is highlighted in the same colour as the node in the query. In this case, the matching node is the node with *t_lemma být* [*to_be*], *functor ACT*, and semantic part of speech *v* (verb), i.e. the word *je* [*are*] in the subordinate clause, which is, as a whole, the *Actor* of the sentence 109c.

[53] http://ufal.mff.cuni.cz/pmltq/doc/pmltq_doc.html

[54] The substitute *t_lemma* is an artificial reconstruction of a lexical value of tectogrammatical nodes in the following cases: newly established nodes that are not copies of other nodes, personal and possessive pronouns, some types of punctuation marks and other symbols, and syntactic negation (see also Chapter 3, Section 3.5.3).

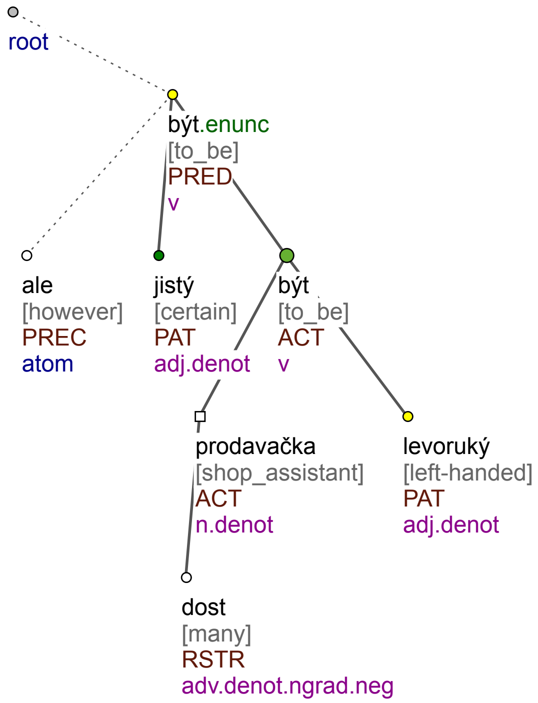


Figure 8.1: The tectogrammatical representation of the resulting sentence 109c for Example 109. The node matching the query is enlarged and highlighted in green (the same colour as the node in the query).

(109c) *Jisté ale je, že je dost levorukých* [*prodavaček*]. (PDT)
It is, however, certain that there are many left-handed [*shop assistants*].

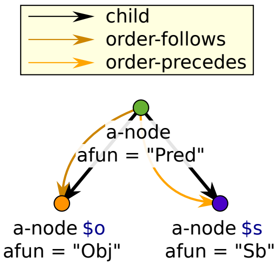
8.1.2 Relations between nodes

Various and multiple relations can be set between pairs of nodes in the query, including 'child', 'descendant', 'sibling', 'same-tree-as', 'same-document-as', but also 'order-follows', etc. The query in Example 110 searches at the analytical (surface syntax) layer of the PDT (see Chapter 6, Section 6.1) for sentences in the OVS (Object–Verb–Subject) order, i.e. for sentences such as *Ondru miluje Marie* [*Ondra is loved by Marie,* lit. *Ondra*$_{\text{Acc\_Obj}}$ *loves Marie*$_{\text{Nom\_Sb}}$]. The query searches for all *Predicates* that directly govern an *Object* and a *Subject*, and specifies that in the left-right order, the *Object* precedes the *Predicate* and the *Subject* is placed after the *Predicate*. These are language phenomena represented at the analytical layer of the PDT, therefore we define analytical nodes (a-nodes) in the query.

(110a) The textual form of the query:

```
a-node
  [ afun = "Pred", order-follows $o, order-precedes $s,
    a-node $o :=
      [ afun = "Obj" ],
    a-node $s :=
      [ afun = "Sb" ] ];
```

(110b) The graphical form of the query:



In the textual version of the query, the first relation between two nodes can be (and usually is) defined by the recursive structure of the query, using square brackets, in this case (110a) with the implicit relation *child*. Additional relations between the same two nodes (e.g. the left-right order) need to be expressed using references to names of the nodes. In this query the *Object* is named *$o*, the *Subject* is named *$s*, and two additional relations are defined using references to these names – *order-follows $o* and *order-precedes $s*. In the graphical version of the query (110b), relations between nodes are expressed by coloured arrows. In the top part of the graphical version of the query, all different types of relations between nodes used in the query are listed, next to arrows in the respective colours.[55]

Figure 8.2 shows one of the results of the query. It is the analytical tree of the sentence 110c. Note that the *Object*, *Predicate* and *Subject* are in the required order.

(110c) *Rozepře prý zinscenoval tisk.* (PDT)
lit. *The_disputes*$_{\text{Acc\_Obj}}$ *allegedly staged the_press*$_{\text{Nom\_Sb}}$.
The disputes were allegedly staged by the press.

[55] The colours of these arrows do not correspond to colours of arrows representing non-dependency relations in the data, i.e. in the result trees, such as textual coreference or discourse relations.

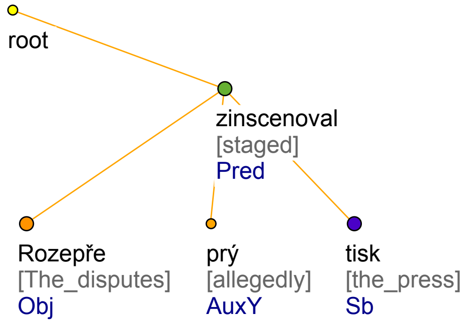


Figure 8.2: The analytical tree of the resulting sentence 110c for Example 110. Nodes matching the query are enlarged and highlighted in colours that match the nodes in the query.

8.1.3 Negative query

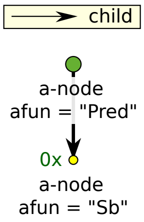
Negation on the level of relations between nodes is a very important part of the query language, as it allows to specify that "we do not wish something in the tree." With this kind of negation, it is possible to search, for example, for sentences without predicates, for predicates without subjects, or for contextually bound expressions without any referential link to the previous context. The PML-TQ uses so-called subqueries to specify how many times a part of the query tree should appear in the result tree (at a given place). "Zero times" then means that it should not be there at all.

The query in Example 111 searches at the analytical layer of the PDT for *Predicates* that do not directly govern a *Subject*, which is technically in the query expressed as a *Predicate* governing a *Subject* "zero times."

(111a) The textual form of the query:

```
a-node
  [ afun = "Pred",
    0x a-node
      [ afun = "Sb" ] ];
```

(111b) The graphical form of the query:



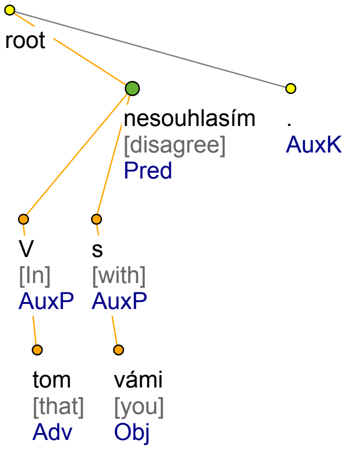


Figure 8.3: The analytical tree of the resulting sentence 111c for Example 111. The node matching the *Predicate* from the query is enlarged and highlighted in the matching colour, unlike, of course, the missing *Subject*.

(111c) *V tom s vámi nesouhlasím.* (PDT)
lit. *In that with you* [*I*] *disagree.*
I do not agree with you on that.

Figure 8.3 shows one of the results of the query. It is the analytical tree of the sentence 111c.[56] As Czech is a pro-drop language, no *Subject* is in this case expressed in the sentence (and neither in the analytical tree as a dependent node of the *Predicate*).[57]

8.1.4 Crossing the layers of annotation

Some queries need to combine information from various layers of annotation, for example to study surface syntax and morphology together or to study relations between the deep and surface representations of the sentence. Example 112 shows an inter-connection of the tectogrammatical layer and the analytical layer in a single query. We are looking at the tectogrammatical layer for a *Predicate* governing an *Actor*, which is not the *Subject* of the sentence in its representation at the analytical layer.

The query defines a t-node with the *functor PRED* and a depending t-node with the *functor ACT*. The connection from the *Actor* to its lexical counterpart at the analytical

[56] The English translations of the Czech word forms in the analytical trees are not a part of the treebank data. The translations have been added to the trees in the figures in this book for easier comprehensibility.

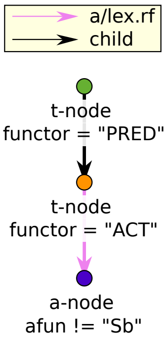
[57] The elided *Subject* of the sentence is reconstructed as an obligatory *Actor* at the tectogrammatical layer, see Chapter 6, Section 6.1.

layer is defined by using the attribute *a/lex.rf*, which is a link to the respective node at the analytical layer;[58] for this a-node, we require that it is not annotated as *Subject* (its analytical function *afun* is not *Sb*).

(112a) The textual form of the query:

```
t-node
  [ functor = "PRED",
    t-node [
      functor = "ACT",
      a/lex.rf a-node
        [ afun != "Sb" ] ] ];
```

(112b) The graphical form of the query:



In one of the results (see the sentence 112c and its tectogrammatical representation in Figure 8.4), the t-nodes from the query match the node representing the passive verb form *je vázána* [*is bound*], together with the dependent node representing the word *dohodami* [*by agreements*]. At the analytical layer, *dohodami* [*by agreements*] is an *Object*.

(112c) *K zákazu je ČR vázána mezinárodními dohodami.* (PDT)
lit. *To the_ban is ČR bound* [*by*] *international agreements.*
The Czech Republic is bound to [*implement*] *the ban by international agreements.*

8.2 Discourse Coherence Phenomena and the PML-TQ

8.2.1 Non-dependency relations

Textual coreference, bridging anaphora and discourse relations (among others) are represented in the data as references from one node (*start node* – the node where the

[58] There is at most one lexical analytical counterpart for each t-node, represented by its identifier in the attribute *a/lex.rf*; for auxiliary analytical counterparts (a-nodes representing prepositions, modal verbs etc.), there is a list of their identifiers in the attribute *a/aux.rf*.

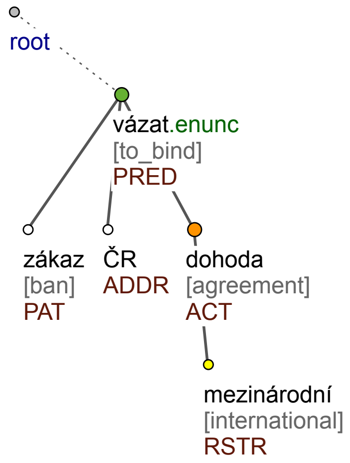


Figure 8.4: The tectogrammatical representation of the resulting sentence 112c for Example 112

relation starts) to another node (*target node* – the node where the relation ends).[59] In the graphical representation of the trees, these relations are depicted as curved arrows connecting the respective two nodes. As several relations of each type may start at a single node and as these relations carry additional information (e.g. the discourse type, scope of the arguments), they are represented in the query language of the PML-TQ as special *member* nodes.

Example 113 shows how to search for a discourse relation of a given type.[60] The query defines two t-nodes connected with a member node that stands for a discourse relation between arguments represented by the two nodes. The required type of the discourse relation can be specified at the member node – in this case it is set to *reason*. The query also specifies that the start and target nodes of the relation are not from the same tree, i.e. it looks for an inter-sentential discourse relation of the type *reason–result*.

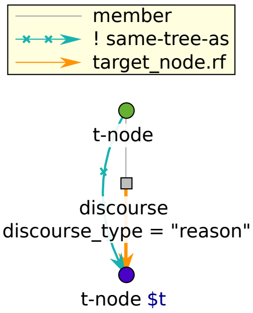
[59] The same technique is used for other relations, like the secondary relation in the verbal complement, or – as shown in Example 112 – the connection between layers of annotation.

[60] Searching for the textual coreference or the bridging anaphora would be very similar, using the respective type of the member node.

(113a) The textual form of the query:

```
t-node
  [ !same-tree-as $t,
    member discourse
      [ discourse_type = "reason",
        target_node.rf t-node $t := [ ] ] ];
```

(113b) The graphical form of the query:



The following two sentences represent one of the results of the query.

(113c) *Neprošel s ní celnicí.* ***Tak***$_{\textbf{reason–result}}$ *si ji pověsil ve své hospodě na stěnu.*[61] (PDT)

lit. *He_did_not_get with it through_customs.* ***So***$_{\textbf{reason–result}}$ REFL *it hung in his pub on the_wall.*

He could not take it through customs. ***So***$_{\textbf{reason–result}}$ *he has hung it on the wall in his pub.*[61]

Figure 8.5 captures the tectogrammatical annotation of these two sentences, along with the discourse relation represented by the thick orange arrow connecting roots of the two respective propositions. Additional relevant information is displayed also in orange at the start node (type of the relation, the connective and the range of the arguments).

8.2.2 Topic-focus articulation and anaphora

The following Example 114 combines some of the techniques described so far to search for a phenomenon studied later in detail in Chapter 13. Specifically, we are interested in non-contrastive contextually bound nodes from which there is no anaphoric reference to the previous context.[62] The query defines a t-node with *tfa* value *t*, from which there is no link of grammatical coreference, no link of textual coreference, and

[61] To give the reader a bit of context, the story is about climbers discussing the perfect prosthesis.

[62] nor any cataphoric or exophoric reference

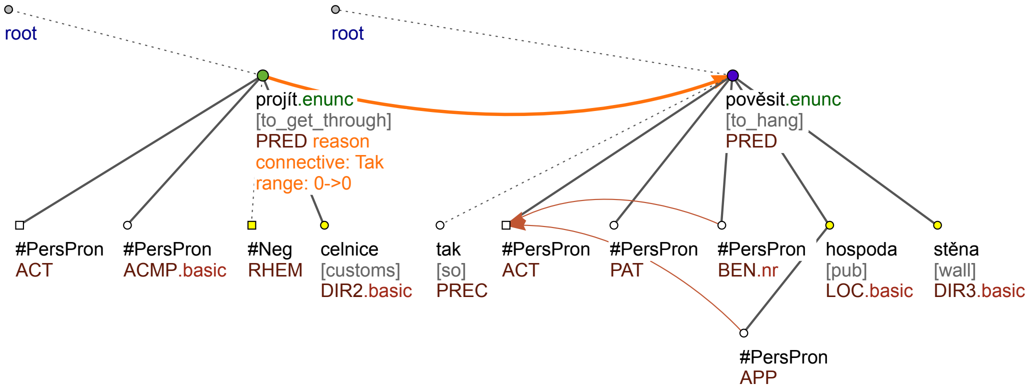


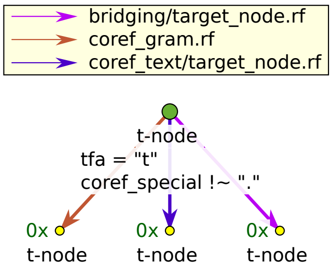
Figure 8.5: The tectogrammatical representation of two resulting sentences 113c for Example 113

no link of bridging anaphora. There may also be no link to an unspecified previous segment and no exophoric link either (both would be captured in the attribute *coref_special* as values *segm* and *exoph*, respectively).

(114a) The textual form of the query:

```
t-node
  [ tfa = "t",
    coref_special !~ ".",
    0x coref_gram.rf t-node [ ],
    0x coref_text/target_node.rf t-node [ ],
    0x bridging/target_node.rf t-node [ ] ];
```

(114b) The graphical form of the query:



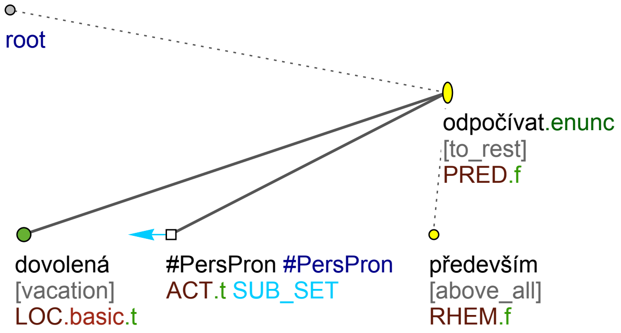


Figure 8.6: The tectogrammatical representation of the resulting sentence 114c for Example 114. Values of the attribute *tfa* are displayed in green next to the *functor*.

(114c) ***Na dovolené*** *chceme především odpočívat.*[63] (PDT)
On vacation*, we want above all to rest.*[63]

Figure 8.6 shows the tectogrammatical representation of the resulting sentence 114c. It is the second sentence of a document and a subheading[64] of the article with an immediately preceding heading 114d. For the reader, the word *dovolená* [*vacation*] is somehow connected to the previous sentence and can be considered contextually bound; however this type of relation is not in any way captured in the PDT annotation.

(114d) *Pojedete do zahraničí s cestovkou?* (PDT)
Will you go abroad with a travel agency?

8.2.3 Output filters

Results of queries can be further processed using *output filters*. Thanks to an output filter, a result of a query does not consist of individual occurrences of the query in the data but instead of a summary of all its occurrences in the searched data, specified by the output filter and presented as a table.[65]

In Example 115, an output filter is added to a simple search. The query defines a single t-node, which is required to be an *Actor* but not a semantic noun (its grammateme *gram/sempos* does not start with *n*) and it does not have a substitute *t_lemma*

[63] a sentence of high significance to the authors of the present book, as the volume was finished during the summer months of 2015

[64] as indicated by the value *heading* in the attribute *discourse_special* and graphically expressed by the oval shape of the node *odpočívat* [*to_rest*]

[65] The result of the output filter can be saved to a textual file in the CSV (comma-separated values) format.

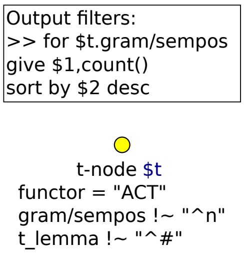
(it does not start with # like e.g. *#PersPron*). The output filter is defined on the last line of the textual form of the query, after the sign '>>'. It states that for each value of the attribute *gram/sempos* found at all nodes matching the query node *$t*, the value of the grammateme (*$1* refers to *$t.gram/sempos*) along with its total count should be listed, and the results should be sorted by the count (referred to by *$2*) in the descending order, i.e. from the most frequent semantic part of speech to the least frequent one.

(115a) The textual form of the query:

```
t-node $t :=
  [ functor = "ACT", gram/sempos !~ "^n", t_lemma !~ "^#" ];

>> for $t.gram/sempos give $1, count() sort by $2 desc
```

(115b) The graphical form of the query:



```
Output filters:
>> for $t.gram/sempos
give $1,count()
sort by $2 desc

        t-node $t
  functor = "ACT"
  gram/sempos !~ "^n"
  t_lemma !~ "^#"
```

Table 8.1 shows the result produced by the query 115 with the output filter.[66] In the left column of the table, the semantic part of speech is listed, in the right column, numbers of occurrences of the respective semantic parts of speech are presented.

8.2.4 Output filters in discourse

Example 116 is similar to Example 113, except that the condition on discourse type has been removed and an output filter has been added. The query searches for all inter-sentential discourse relations (of any discourse type) and the output filter (very similar to the output filter from the previous example) summarizes the results in a distribution of discourse types in the relations.

[66] Numbers in tables in this chapter correspond to a search in 9/10 of the Prague Dependency Treebank, available at the public search server. The remaining 1/10 of the data serve as test development data and therefore are not accessible this way.

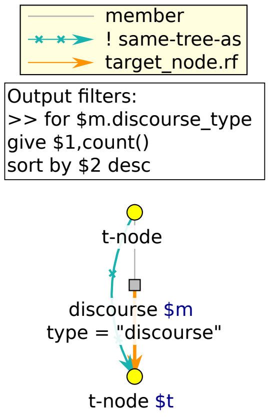
gram/sempos	Number of occurrences
v	3,028
adj.quant.grad	75
adv.denot.grad.neg	31
adj.denot	28
adv.pron.def	24
adj.quant.def	15
adv.denot.ngrad.nneg	4
adv.denot.ngrad.neg	2
adj.pron.def.demon	2
adj.quant.indef	1

Table 8.1: The resulting table for Example 115. The value *v* stands for a semantic verb, values starting with *adj* are semantic adjectives, and values starting with *adv* semantic adverbs. Semantic adjectives and adverbs are further subcategorized, for example *adj.quant.grad* means a gradable quantificational semantic adjective (adjectives such as *mnoho* [*many*]). Headings of the columns are not a part of the query result.

(116a) The textual form of the query:

```
t-node
  [ !same-tree-as $t,
    member discourse $m :=
      [ type = "discourse",
        target_node.rf t-node $t := [ ] ] ];
>> for $m.discourse_type give $1, count() sort by $2 desc
```

(116b) The graphical form of the query:



Output filters:
>> for $m.discourse_type
give $1,count()
sort by $2 desc

Discourse type	Number of occurrences
opp	1,601
conj	1,255
reason	902
confr	272
conc	236
preced	215
grad	184
restr	149
explicat	121
corr	110
...	

Table 8.2: First 10 rows in the resulting table for the Example 116 (headings of the columns are not a part of the query result).

Table 8.2 shows the result produced by the query 116 with the output filter.

A more advanced output filter is used in Example 117, which shows how output filters (or, in other words, lines of an output filter) can be put one after another – the second line of an output filter is applied to the output of the first one, etc. The query 117 searches for all discourse relations in the data and the output filter summarizes the results in a distribution of all, intra- and inter-sentential usages of connectives in the relations (regardless of the discourse types they represent), both in total counts and percentages.

The first line of the output filter produces a table[67] with a row for each discourse relation matching the query, consisting of three values (i.e. the output table has three columns) – a lowercased connective along with the information whether the given relation is intra-sentential (the condition on tree numbers of the start and target nodes produces *1* in the second column) or inter-sentential (*1* in the third column). The second line of the output filter (applied to the output of the previous line) adds up the intra-sentential and inter-sentential occurrences of relations for the various connectives (*$1*, *$2* and *$3* refer to the respective columns in the result of the previous line of the output filter), and for the purpose of counting percentages of these numbers adds the fourth column – a total number of occurrences of the given connective in the relations (intra- and inter-sentential ones together). The third line (applied to the output of the second line) counts the percentages and formats the output (by reorganizing the order of columns and by adding parentheses and percentage marks).

[67] a temporary table, further processed by the subsequent lines of the output filter

(117a) The textual form of the query:

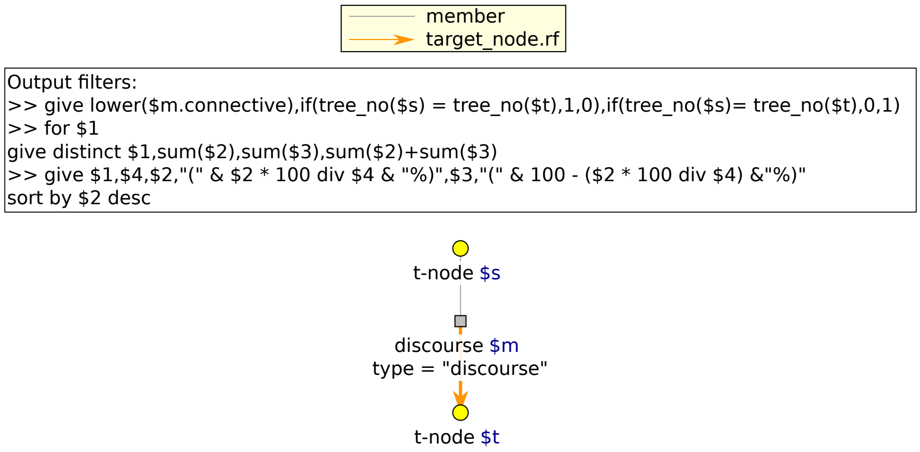
```
t-node $s :=
  [ member discourse $m :=
    [ type = "discourse", target_node.rf t-node $t := [ ] ] ];

>> give lower($m.connective), if(tree_no($s) = tree_no($t),1,0), if(tree_no($s)
= tree_no($t),0,1)

>> for $1 give distinct $1, sum($2), sum($3), sum($2)+sum($3)

>> give $1,$4,$2,"(" & $2 * 100 div $4 & "%)",$3,"(" & 100 - ($2 * 100 div $4) &
"%)" sort by $2 desc
```

(117b) The graphical form of the query:



```
Output filters:
>> give lower($m.connective),if(tree_no($s) = tree_no($t),1,0),if(tree_no($s)= tree_no($t),0,1)
>> for $1
give distinct $1,sum($2),sum($3),sum($2)+sum($3)
>> give $1,$4,$2,"(" & $2 * 100 div $4 & "%)",$3,"(" & 100 - ($2 * 100 div $4) &"%)"
sort by $2 desc
```

Table 8.3 shows the result produced by the query 117 with the output filter.

8.3 Hands on the Data

The last section of this chapter is dedicated to a highly technical matter – how to get the data – and can be safely skipped if the reader is not interested in either downloading the PDT or searching in it using the PML-TQ query language described above.

8.3.1 Data to download

The PDT 3.0 is freely available from the public repository of the Lindat/Clarin project[68] under the Creative Commons Licence.[69] The corpus data are stored in the Prague Markup Language format (PML; http://ufal.mff.cuni.cz/jazz/PML/), which is an abstract XML-based format designed to capture complex annotations of language data,

[68] http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3

[69] Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported Licence (CC BY-NC-SA 3.0)

Connective	Total	Intra-sentential	(%)	Inter-sentential	(%)
a [*and*]	5,128	4,815	(93%)	313	(7%)
však [*however*]	1,356	236	(17%)	1,120	(83%)
ale [*but*]	1,134	758	(66%)	376	(34%)
když [*when*]	478	478	(100%)	0	(0%)
protože [*because*]	469	463	(98%)	6	(2%)
totiž [*actually, in fact*]	405	20	(4%)	385	(96%)
:	353	310	(87%)	43	(13%)
pokud [*if*]	342	342	(100%)	0	(0%)
proto [*therefore*]	339	32	(9%)	307	(91%)
aby [*to*]	276	275	(99%)	1	(1%)
tedy [*therefore*]	269	30	(11%)	239	(89%)
pak [*then*]	257	66	(25%)	191	(75%)
ovšem [*however*]	257	57	(22%)	200	(78%)
li [*if*]	227	227	(100%)	0	(0%)
také [*also*]	208	7	(3%)	201	(97%)
neboť [*because*]	196	196	(100%)	0	(0%)
–	194	193	(99%)	1	(1%)
zatímco [*while*]	175	174	(99%)	1	(1%)
nebo [*or*]	169	150	(88%)	19	(12%)
navíc [*moreover*]	169	24	(14%)	145	(86%)
...					

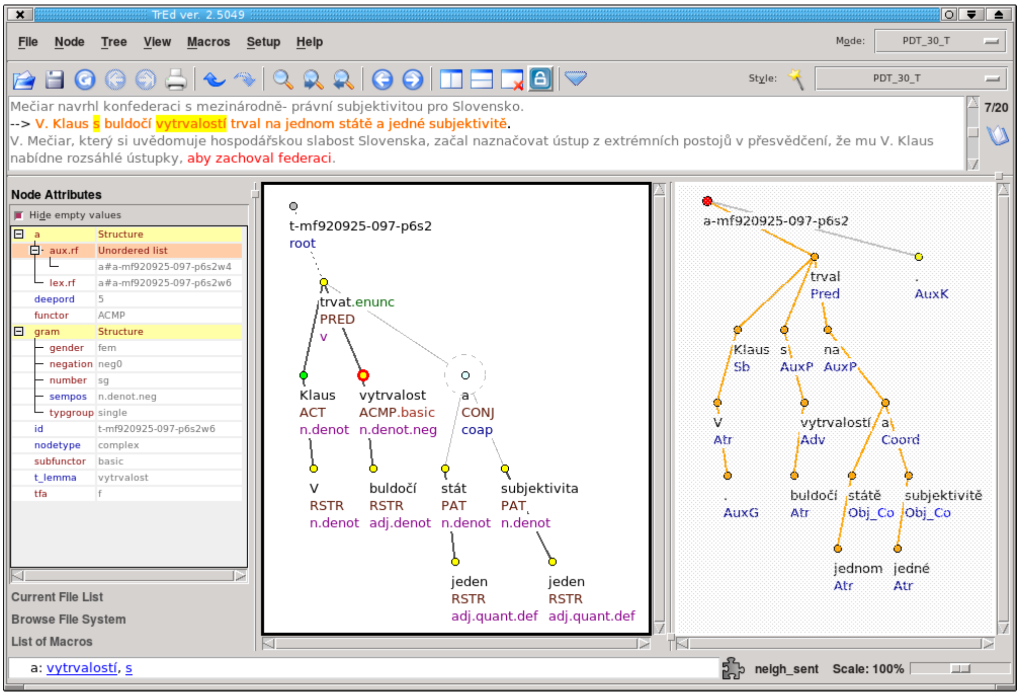
Table 8.3: First 20 rows in the resulting table for Example 117 (headings of the columns and the translations are not a part of the query result).

particularly treebanks. Although the format is text-based, it is difficult to read in the raw form, more so at the higher layers of annotation. The tree editor TrEd (see below) should be used for more convenient viewing.

Tree editor TrEd

Tree editor TrEd (Pajas and Štěpánek, 2008) is a primary tool for browsing (and editing) the PDT 3.0 data. It can be freely downloaded (for various platforms, including MS Windows and Linux) from its home web page[70] under the GPL – The General Public Licence. After the installation of the editor itself, an extension for the PDT 3.0

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



Mečiar navrhl konfederaci s mezinárodně- právní subjektivitou pro Slovensko.
--> V. Klaus s buldočí vytrvalostí trval na jednom státě a jedné subjektivitě.
V. Mečiar, který si uvědomuje hospodářskou slabost Slovenska, začal naznačovat ústup z extrémních postojů v přesvědčení, že mu V. Klaus nabídne rozsáhlé ústupky, aby zachoval federaci.

Node Attributes

Hide empty values

a	Structure
aux.rf	Unordered list
	a#a-mf920925-097-p6s2w4
lex.rf	a#a-mf920925-097-p6s2w6
deeord	5
functor	ACMP
gram	Structure
gender	fem
negation	neg0
number	sg
sempos	n.denot.neg
typgroup	single
id	t-mf920925-097-p6s2w6
nodetype	complex
subfunctor	basic
t_lemma	vytrvalost
tfa	f

Figure 8.7: The tectogrammatical and analytical representations of the sentence *V. Klaus s buldočí vytrvalostí trval na jednom státě a jedné subjektivitě* [*V. Klaus with bulldog persistence insisted on a single state and a single* [*legal*] *subjectivity*], displayed in TrEd.

support needs to be installed as well, which can be done from the TrEd menu using Setup → Manage Extensions → Get New Extensions.[71]

Afterwards, any document from the PDT can be opened in TrEd. Annotation of each document is stored in four files corresponding to the word layer (file with the suffix *.w*), the morphological layer (file with the suffix *.m*), the analytical layer (file with the suffix *.a*), and the tectogrammatical layer (file with the suffix *.t*).[72] Any of these files except for the files of the word layer can be opened in TrEd; for displaying the tectogrammatical annotation of a given document, the respective file with the suffix *.t* needs to be opened, the file with the suffix *.a* for the analytical annotation, and the file with the suffix *.m* for the morphological annotation.[73] Figure 8.7 shows

[71] A more detailed description of the installation can be found in the documentation for the PDT 3.0 at http://ufal.mff.cuni.cz/pdt3.0/data.

[72] All files are compressed by gzip, which means that their suffixes are in fact *.w.gz*, *.m.gz*, *.a.gz* and *.t.gz*.

[73] TrEd is a tree editor. Therefore, it cannot be directly used to open files of the word layer. However, thanks to the interlinking of the layers, the information at the w-layer is accessible from the higher layers. Files

the graphical interface of TrEd, which consists of several sections for displaying the data, namely the textual area on top (it displays the sentence in its context) and – in this case – two areas for the tectogrammatical and analytical representations of the given sentence (in the middle and on the right). Values of all non-empty attributes for a selected node are displayed in the left panel.

8.3.2 Data for searching

The PDT 3.0 data can be searched on a public search server, i.e. without a prior download, using the PML-TQ – the query language described in the previous sections.

There are two ways of accessing the search server for the PDT. The first method uses the tree editor TrEd along with a PML-TQ extension.[74] The second method accesses the server using a web browser; the server for PDT 3.0 data is publicly available at the LINDAT/CLARIN portal.[75] The web-based access has several limitations, namely less variability in displaying the results and the necessity to create the query in the textual form. TrEd, on the other hand, needs to be installed first (along with the PML-TQ extension), but it does offer a better user interface, the query can be created graphically and the graphical representation of the results can be adapted to the user's needs.

We have introduced two possible ways how to get one's hands on the data of the Prague Dependency Treebank – downloading the corpus or searching in it on-line. As such, the data are open to experiments and easily accessible for studying many language phenomena. It is very important that various levels of annotation/language description are annotated separately but can be used together, even in a single search query. It opens the possibility for researchers to study – if we stay on the topic of this book – the interplay among morphology, the syntactic structure of the sentence, discourse relations, anaphora, and the topic–focus articulation. Several such case studies are presented in the subsequent part of the book.

of the morphological layer can be opened in TrEd, because, technically, each sentence is at this layer represented as a sequence of nodes corresponding to the words of the sentence, with a single technical root as their common parent. This solution would not be practical for the w-layer, as the data at the w-layer are not segmented into individual sentences.

[74] See the on-line documentation of the PML-Tree Query for instructions on how to install the extension: http://ufal.mff.cuni.cz/pmltq/.

[75] https://lindat.mff.cuni.cz/services/pmltq/

Case Studies

9

Relation of Discourse Analysis to Syntax

The following case study focuses on the relations between discourse-level analysis and syntactic analysis of the text. When we were considering the possibilities of discourse analysis (i.e. annotating discourse relations)[76] there were basically two ways of dealing with it in Czech – either to use plain texts and mark discourse relations directly in these texts or to add discourse analysis as new information to the existing morphological and syntactic analyses in the Prague Dependency Treebank (for a description of this corpus see Chapter 6).

While the majority of approaches to discourse-level analysis use raw written documents as annotation basis,[77] in discourse-level analysis of Czech we decided to annotate discourse relations directly on syntactically annotated layer (on the top of tectogrammatical dependency trees), making a basic assumption that some syntactic features of a sentence analysis on this layer correspond to certain discourse-level features. Moreover, we are convinced that multilayer annotation provided for the same data can reveal new insights into text structure (and language use). Issues discussed in this case study resulted from our decision to annotate discourse relations directly on the tectogrammatical layer, since during the whole process of data development we encountered a number of new challenges and questions on the relation between syntactic and discourse-level analyses.

For this case study, we want to approach this relation in two steps, from two points of view. We want to

(i) describe and evaluate all features of the tectogrammatical layer which were used during discourse annotation and evaluate their role during annotation of discourse relations,

(ii) show what we can learn about discourse relations using syntactic annotation.

Accordingly, the chapter is divided into two parts – the features of syntactic analysis used during discourse annotation are described in Section 9.1, observations of discourse relations from the syntactic point of view are presented in Section 9.2. In both sections, the study provides partly new insight and comments, and partly summarizes findings made in Jínová, Mírovský and Poláková (2012b); Mírovský, Jínová and Poláková (2012); Jínová, Poláková and Mírovský (2014).

[76] For details see Chapter 2.

[77] Cf. e.g. the RST Discourse Treebank (Carlson, Okurowski and Marcu, 2002), the Penn Discourse Treebank (Prasad et al., 2008) for English, the Potsdam Commentary Corpus (Stede, 2004) for German.

9 RELATION OF DISCOURSE ANALYSIS TO SYNTAX

In this chapter, we demonstrate that syntactic analysis[78] approaches discourse structure from a different angle than discourse-level analysis. In brief, syntactic analysis deals with form to the same extent as with meaning, while discourse-level analysis focuses on meaning disregarding form.

9.1 Features of Syntactic Analysis Used for Discourse-Level Analysis

One of the most important features of syntactic analysis in the PDT with regard to the discourse-level analysis is the representation of the syntactic structure of sentences.[79] Each sentence is represented by a separate dependency tree, all relations between clauses are captured and semantically interpreted. The difference between dependency of clauses and coordination between them is clear at first glance. Some types of relations between clauses are discourse-relevant. Parentheses are also treated as parts of the tree and can be relevant for analysis of discourse structure. Besides, connective devices are part of the representation, too.

Example 118 illustrates all the mentioned syntactic features in one complex sentence.

(118) *Například:* (I) *jestliže matka nechtěla dítě a* (II) *dítě se jí narodilo proti její vůli,* (III) *vyvíjelo se nepříznivě a* (IV) *je vysoká pravděpodobnost, a* (V) *my chceme vědět jaká,* (VI) *že i toto dítě se v budoucnu bude chovat ke svému dítěti podobně.* (PDT)

For example:[80] (I) *if a mother did not want to have a child and* (II) *the child was born against her will,* (III) *its development was unhealthy and* (IV) *there is a high probability* – (V) *and we want to know how high* – (VI) *that in the future this child will also behave similarly towards its own child.*

The first two clauses (I) and (II) are coordinated[81] and together dependent on coordination of the clauses (III) and (IV). Coordinations are interpreted from the syntactico-semantic point of view as *Conjunction* (label *CONJ* occurs between both dependent and governing clauses, cf. Figure 9.1). The clauses (I) and (II) are syntactico-semantically interpreted as *Condition* (label *COND* at the highest node of both sentences). The clause (V) is connected with the fourth clause only loosely and the former comments the content of the latter – thus it is a *Parenthesis* (label *PAR* at the highest node of

[78] Or at least syntactic analysis in the Functional Generative Description perspective (Sgall, Hajičová and Panevová, 1986) which is the theoretical basis for syntactic annotation of the PDT (cf. Chapters 1 and 6).

[79] For a general overview see Chapter 6.

[80] The sentence explains the meaning of the notion of pathology of the third generation that appeared in a previous context.

[81] The terms *coordination* and *subordination* are used in this chapter for two different types of syntactic structure in a complex sentence. Basically, a coordinate structure contains two (or more) formally independent clauses – the whole structure can be split into separate sentences. A subordinate structure, on the contrary, represents one syntactic whole – a subordinate clause cannot stand alone, it is closely connected to its governing clause.

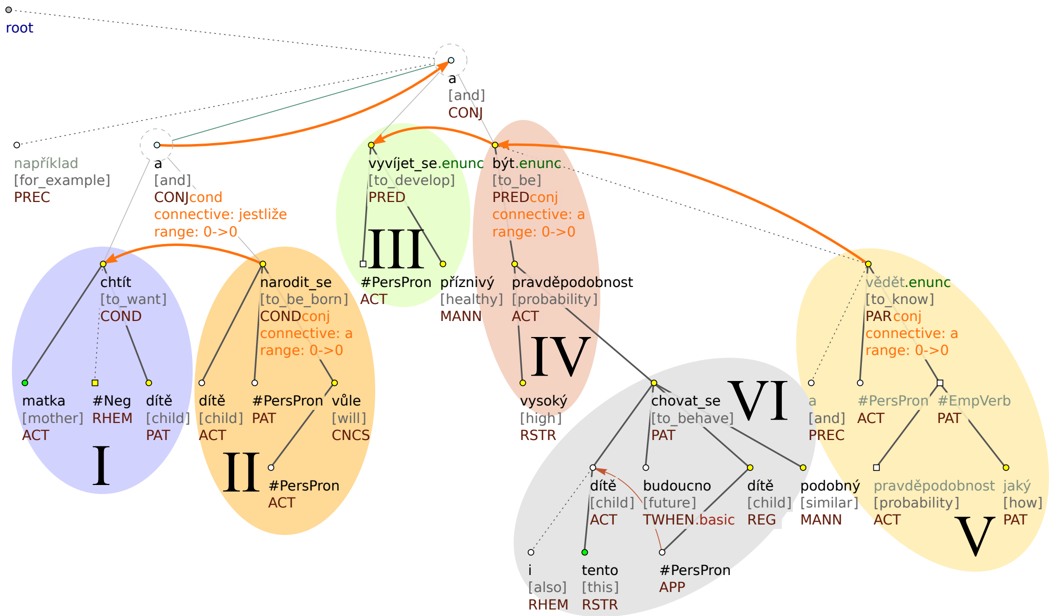


Figure 9.1: The tectogrammatical representation of Example 118

the structure signals this function). The clause (VI), depending on the clause (IV), is interpreted as a part of the valency frame of the predicate of the governing clause (*je pravděpodobnost* [(*there*) *is a probability*]) and thus has the label *PAT* (*Patiens*).

Apart from the syntactico-semantic representation of the compound sentence on the tectogrammatical layer, Example 118 illustrates the representation of discourse relations as well.[82] Four of them are represented by thick orange arrows – three discourse *conjunctions* (abbreviation *conj* in Figure 9.1) with connective *a* [*and*] and one discourse *condition* (abbreviation *cond* in Figure 9.1) with connective *jestliže* [*if*]. Discourse *conjunctions* hold between clauses:

- (I) *a mother did not want to have a child* ***and*** (II) *the child was born against her will*
- (III) *its development was unhealthy* ***and*** (IV) *there is a high probability* (VI) *that in the future also this child will behave similarly towards its own child*
- (IV) *there is a high probability* (VI) *that in the future this child will also behave similarly towards its own child* ***and*** (V) *we want to know how high*

[82] For a basic description of the tectogrammatical representation see Chapter 6.

Discourse *condition* holds between clauses *if a mother did not want to have a child, the child was born against her will* and clauses *its development was unhealthy, there is a high probability that in the future also this child will behave similarly towards its own child*.

A scope of arguments of all these discourse relations is indicated by the tree structure (the arguments equal subtrees, see Section 9.1.2), connectives are either part of the tectogrammatical representation (if they occur in coordinate structures) or they are part of the surface syntactic layer which is interconnected with the tectogrammatical layer (this is the case for connectives in dependent clauses, see Section 9.1.3).

The structure of Section 9.1 is quite complex, so, for better orientation in the text, we provide an outline here.

Given the features of syntactic analysis on the tectogrammatical layer exploited during discourse annotation, Section 9.1 is divided into three main parts – Section 9.1.1 describes utilization of syntactico-semantic labels (*functors*) in discourse annotation, Section 9.1.2 discusses topics connected with the scope of discourse arguments and finally Section 9.1.3 focuses on connectives of discourse relations.

Section 9.1.1 comprises four parts, first of them is further divided into four minor subsections, the structure of Section 9.1.1 is thus as follows:

- Syntactic analysis captures form-based characteristics of structure
  - syntactic structures with apposition
  - relative clauses
  - clauses with connective *s tím, že* [*along with*]
  - structures treated according to the prototypical meaning of their conjunctions
- Syntactic analysis captures general semantic characteristics
- Syntactic analysis captures semantic relations between clauses from a different perspective than discourse-level analysis
- Syntactic analysis directly transferable to discourse-level analysis

Section 9.1.2 discusses the following topics:

- Effective subtree as a discourse argument
- Coordination resolution
- Ellipsis resolution

And finally Section 9.1.3 comprises these subsections:

- Expression referring to preceding context
- Rhematizing particles
- Automatic detection of discourse connectives
- Expression *což* [*which*]

The structure of all parts is discussed in more detail below.

9.1.1 Syntactico-semantic labels for relations between clauses

As mentioned in Chapter 2, the annotation of discourse relations was carried out in two phases – first, annotators went through all texts marking possible discourse connectives and then annotated the relations[83] on the tectogrammatical layer of the PDT manually (therefore we will call this phase *manual*). The second phase was almost completely automatic, it consisted of an automatic procedure that extracted mostly tectogrammatical syntactic features and used them directly for the annotation of intra-sentential discourse relations (therefore, this phase will be called *automatic*). It should be noted, however, that the automatic phase took place only after the manual one – annotators specified places where the automatic extraction was possible by leaving them unmarked, without manual annotation.

The remaining part of Section 9.1.1 covers the following issues from both phases of the annotation: First, three subsections describe correspondence between tectogrammatical syntactico-semantic labels (functors) and discourse types during the manual phase of annotation, the last subsection deals with the utilization of these labels for automatic extraction of discourse relations.

Let us start with phenomena connected with syntactico-semantic labels encountered during the manual phase of discourse relations annotation. During this phase, only such relations were marked for which syntactic analysis on the tectogrammatical layer does not offer sufficient information from the discourse-analysis point of view. Having completed data annotation, we can see that mismatches between syntactic and discourse analysis created three groups:

- cases where syntactic analysis conveys form-based characteristics rather than meaning,
- cases where discourse-level analysis needs to capture finer semantic distinctions than those offered by the syntactic analysis, and finally
- cases where annotators of discourse relations decided to newly annotate syntactico-semantic relations present in syntactic analysis from a discourse perspective.

All three types are described in detail below.

Syntactic analysis captures form-based characteristics of structure

Syntactic phenomena relevant for discourse-level analysis, but at the same time represented formally rather than semantically, were found to be:

- structures with apposition (syntactic label *APPS*) between clauses (i.e. structures with clauses conveying more of less the same meaning),[84]

[83] Mostly inter-sentential but also some intra-sentential, see Chapter 2, Section 2.6.4.

[84] For relation between non-clauses, this label is used in cases like *Antonín Dvořák,*APPS *a Czech composer* or *these special topics,*APPS *e.g. literature of Croatia in the last decade*.

- structures with some relative clauses (syntactic label *RSTR*),[85]
- structures with dependent clauses with connective *s tím, že* (roughly [*along with*] or [*saying also that*], lit. [*with that that*]) labelled *ACMP*[86] and
- structures with dependent clauses whose characteristics are assigned to them on the basis of their formal features only (e.g. according to the connective).

For all these cases, manual discourse annotation was necessary as discourse-level meaning can be assigned to them only according to the content of both clauses. We will now describe the basic linguistic features for all these structures and we will also provide some basic frequency characteristics to demonstrate how common these phenomena are in our data.

From the beginning of discourse annotation we assume that **syntactic structures with apposition** most often convey the meaning corresponding at discourse level to the notion of adordination, i.e. to semantic types *specification*, *instantiation*, *generalization* and *equivalence* which represent relations between parts of text conveying the same or at least partly the same meaning. Keeping this assumption in mind the syntactic structures with apposition were annotated as discourse-level *specification* in 60% of the approximately 180 cases;[87] the total amount of types corresponding to discourse-level adordination is then 72%. However, other discourse relations occur in connection with this structure too, namely *conjunction* (20% of all cases), *reason–result*, *gradation*, *correction*, *restrictive opposition* and *opposition*.

Example 119 illustrates the most common situation:[88] a compound sentence with a syntactic apposition between clauses and the discourse relation of *specification* between them (the first clause conveys a general description of a situation, the rest of the sentence describes the same situation in detail). Example 120 documents a similar case for the discourse-level relation of *reason–result* (the first clause contains a claim, the rest of the sentence a reason/motivation for the claim). In both cases, the connective device is a colon.

(119) <Arg1: *Jako by pouze vzájemně prohodil obvyklé mužské a ženské role*>***:*****specification** <Arg2: *milenec jedné z dívek je zastřelen zezadu před jejíma očima a zahyne tedy obdobně, jako ve westernech zpravidla umírají partnerky mužských hrdinů.*> (PDT)

[85] The *RSTR* is a label for a free modification further restricting or specifying a governing noun, so e.g. for modifications as *a small.*RSTR *city* or *a company operating.*RSTR *in the Czech Republic*.

[86] The *ACMP* is an abbreviation for *Accompaniment*, it is a label for such an adjunct which expresses manner by specifying a circumstance that accompanies the event or entity modified by the adjunct, so e.g. *to walk with a dog.*ACMP or *all costs including project preparation.*ACMP. For all labels see Mikulová et al. (2006) or http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/html/index.html.

[87] All numbers reported in this chapter refer to the training and development test parts of the data, i.e. 43,955 sentences (approximately 9/10 of the treebank). The *evaluation test* part of the data thus remains unobserved.

[88] In all the following examples, connectives are printed in bold, and, where relevant, the type of discourse relation is added to the connective. Other relevant phenomena are underlined, where necessary.

<Arg1: *As if he simply switched conventional male and female roles>*:**specification** <Arg2: *the lover of one of the girls is shot in the back right in front of her and thus dies in the way the girlfriends of heroes in westerns usually do.>*

(120) <Arg1: *Nebylo to však možné>*:**reason–result** <Arg2: *došlo by tím k fatálnímu zásahu do všech stávajících právních vztahů, přestaly by platit působnosti státních orgánů a stát by nemohl existovat.>* (PDT)

<Arg1: *However, it was not possible>*:**reason–result** <Arg2: *that would be a critical intervention into all existing legal relations, the powers of state authorities would be invalidated and the state could not exist.>*

As for **relative clauses**, only those are considered to be relevant for discourse-level analysis that convey an independent meaning rather than a mere characteristics of the governing noun. These clauses are often called *false relative clauses* in the Czech linguistic tradition.[89] When we searched for a way to test for distinguishing discourse relevant (i.e. those with independent meaning) and non-relevant cases, possible reformulation with some common connective as *a* [*and*], *ale* [*but*] or *když* [*when*] turned to be useful. However, there is a number of cases where it is difficult to make a distinction between discourse-relevant and non-relevant relative clauses. Therefore, we decided to annotate only such relative clauses that are connected to their governing clause also by some other connective beside the relative pronoun or adverb (cf. simultaneous occurrence of relative pronoun *which* and connective *however* in Example 121).

Examples 121–123 illustrate the relative clauses which were newly annotated according to their discourse semantics. In the first one, the most frequent type of discourse relation in this structure – *opposition* – is present. The second one exemplifies another frequent case – the *asynchrony* relation, and the third one gives the only instance of a relative clause with a *conditional* discourse meaning in our data.[90]

(121) *V našem případě by se mělo jednat o zákon o zdraví lidu,* ***který***relat. pron. *se* ***však***opposition *nedá očekávat dříve než za dva roky.* (PDT)

In our case it should be a public health law, ***which***relat. pron., ***however***opposition, *cannot be expected earlier than in two years.*

(122) *Přáli si okamžité jednání s vedením celnice, policií a okresním úřadem, na* ***kterém***relat. pron. ***pak***asynchrony *předložili devět požadavků.* (PDT)

They requested an immediate meeting with the senior officials from customs, police and the district office, during ***which***relat. pron. *they* ***then***asynchrony *submitted nine demands.*

[89] For the summary of linguistic approaches to these structures see Daneš (2009), for an overall characteristics of relative clauses in Czech cf. Fried (2011).

[90] In the last case, the connective *ovšem*, which often signals *opposition*, can be replaced by the typical conditional connective *if*, therefore the relation is annotated as *condition*.

(123) *Téměř stejně ceněný je obraz v nepůvodním stavu, s restaurátorskými zásahy,* ***které***$_{\text{relat. pron.}}$ *jsou* ***ovšem***$_{\textbf{condition}}$ *kvalitní.* (PDT)

Almost equally valued is a painting in unoriginal state, with restoration interventions, ***which***$_{\text{relat. pron.}}$ *have to be,* ***however***$_{\textbf{condition}}$*, of good quality.*

There are approximately 100 relative clauses relevant for discourse-level analysis in our data – the most common semantic discourse type annotated for these structures is *opposition* (approx. 40 cases), succeeded by *conjunction* (approx. 20 cases), *asynchrony* (approx. 20 cases), *restrictive opposition*, *reason–result* and marginally some other types (i.e. *confrontation*, *correction*, *gradation*, *disjunctive alternative*, *condition*, *pragmatic contrast* and *explication*). Contrary to our assumption, temporal *asynchrony* relation is not the most frequent discourse type for these clauses, although it is treated as the most prototypical case of the so-called *false relative clauses* in Czech grammar handbooks.[91]

The third type of structure that is treated in syntactic analysis primarily according to its form (and as such rather underspecified for discourse-level analysis) is represented by dependent **clauses with connective *s tím, že*** (roughly [*along with*] or [*saying also that*], lit. [*with that that*]). This connective lacks any precise meaning,[92] in other words, it can serve as a connective for many discourse relations. The type of discourse relation can be inferred from the context only. The expression *s tím, že* [*along with, saying also that*] serves as a discourse connective in the PDT data approximately in 30 cases, two most frequent types of relation are *conjunction* and *reason–result*, but we found also some cases of *specification*, *condition*, *explication*, *synchrony* and *asynchrony*. Contexts for the two most common types (*conjunction* and *reason–result*) are given in Examples 124 and 125, respectively.

(124) *Tyto komplexy by pak měly být navrženy k privatizaci* ***s tím, že***$_{\textbf{conjunction}}$ *v nich bude stanoven podíl státu.* (PDT)

These industrial complexes should be then offered for privatization ***and***$_{\textbf{conjunction}}$ (lit. ***with that that***) *the share of the state will be specified.*

(125) *Blažek odmítl návrh strany na své vystoupení komentovat* ***s tím, že***$_{\textbf{reason–result}}$ *je to věcí vedení strany.* (PDT)

Blažek refused to comment on his party´s proposal for his resignation, ***saying that***$_{\textbf{reason–result}}$ (lit. ***with that that***) *it is a matter for the party leadership.*

Finally, some differences between syntactic and discourse-level analyses arise from the fact that all structures are syntactico-semantically treated according to the **prototypical meaning of their conjunctions**. For example, the prototypical meaning

[91] e.g. Grepl and Karlík (1998), Daneš et al. (1987)

[92] Therefore, syntactically it is treated according to the form only, as *Accompaniment*.

conveyed by the conjunction *jestliže – tak* [*if – then*] is condition, therefore the dependent clause in Example 126 is syntactically treated as *condition* on the tectogrammatical layer, while from the discourse-level perspective it represents a *confrontation* relation. This situation was found mostly for conjunctions *když* [*when*], *jestliže* [*if*], *-li* [*if*], *kdyby* [*if*] and *aby* [*in order to*]. They are typical for syntactically temporal (conjunction *když*), purpose (conjunction *aby*) and conditional (all of them apart from *aby*) relations, but from the viewpoint of discourse analysis, in specific contexts they express other meanings: most frequently *confrontation*, *conjunction* (illustrated by Examples 127 and 128) and *specification* (see Example 129).

(126) ***Jestliže***$_{\textbf{confrontation}}$ *v roce 1993 jich bylo 8650, což je vytížení kapacity lázní asi na 65 až 70 procent,* ***tak*** *v letošním roce by jich mělo být již 9745.* (PDT)

While$_{\textbf{confrontation}}$ (lit. ***if***) *there were 8,650 of them*[93] *in 1993, which represents the utilization of the spa capacity to about 65 to 70 percent,* (lit. ***then***) *this year there should already be 9,745 of them.*

(127) *Pod jmény slavných finských spisovatelů publikovala své vlastní texty,* ***když***$_{\textbf{conjunction}}$ *z původních děl přejala pouze vlastní jména.* (PDT)

Under the names of famous Finnish writers, she published her own texts, ***and***$_{\textbf{conjunction}}$ (lit. ***when***) *she took over only proper names from the original works.*

(128) *Obchodní vztahy mezi Českou republikou a Dánskem se začaly výrazněji rozvíjet až po roce 1990,* ***když***$_{\textbf{conjunction}}$ *nejvyššího rozmachu dosáhly v roce 1992.* (PDT)

Trade relations between the Czech Republic and Denmark began to develop significantly after 1990, ***when***$_{\textbf{conjunction}}$ *they reached an all-time high in 1992.*

(129) *Sparta v dramatickém utkání remizovala v Chebu,* ***když***$_{\textbf{specification}}$ *brankář Kouba kryl i penaltu chebského Bielika.* (PDT)

Sparta drew in a dramatic match in Cheb, ***when***$_{\textbf{specification}}$ *the goalkeeper Kouba also parried a penalty by Cheb´s Bielik.*

We encountered approximately 80 such cases in our data. Some of them are described in the literature (especially such cases as Example 126, i.e. discourse *confrontation* in syntactic structure with conditional conjunction), but some of them represent so far undocumented structures (Examples 127–129). There are also stylistic differences between these structures. While compound sentences in Examples 126 and 129 are quite common in Czech, cases like Examples 127 and 128 represent stylistic ineptitude and in a well-formulated text their connectives should be replaced by conjunction *and*.

[93] The text discusses the number of visitors at a spa.

Discourse types corresponding to syntactic *Adversative* relation	Number of occurrences
opposition	1,093
correction	174
concession	80
confrontation	59
restrictive opposition	48
pragmatic opposition	14
gradation	3

Table 9.1: Discourse relations from CONTRAST class corresponding to the syntactic *Adversative* relation (label *ADVS*) between clauses on the tectogrammatical layer

Syntactic analysis captures general semantic characteristics

After discussing syntactic structures captured according to their form, let us discuss the second type of constructions where it was necessary to carry out manual discourse annotation. This type is represented by coordination of clauses syntactically labelled as *Adversative* relation (syntactic label *ADVS*). From the discourse-level perspective, this label corresponds to several finer discourse semantic types from the CONTRAST class, namely to *opposition*, *restrictive opposition*, *correction*, *confrontation*, *concession*, *pragmatic opposition* and *gradation*. *Opposition* is the most frequent discourse relation occurring in the syntactically *Adversative* structures (it represents slightly more than 70% of approximately 1,500 occurrences of *Adversative* relations between clauses in our data), therefore, it was left aside for the second phase – automatic extraction of discourse relations. Other discourse types were annotated manually during the first phase – they represent approximately 25% of all cases (exact numbers for all relations are shown in Table 9.1).

Examples 130 and 131 illustrate the cases where the syntactic *Adversative* relation between clauses does not correspond in discourse-level analysis to *opposition*, but rather to *correction* and *restrictive opposition*, respectively, which are semantically more specific than *opposition*. *Correction* is a relation in which the content of the second argument corrects or replaces the content of the first argument. Typically, the first argument contains negation. *Restrictive opposition* is a relation in which the validity of the first argument is limited by the content of the second argument or the second argument expresses an exception to the first (see Poláková et al., 2012a, pp. 44–47).

(130) *Sestupná tendence porodnosti přitom není jevem poslední doby,* ***ale***$_{\textbf{correction}}$ *výrazně se prohlubuje od poloviny 60. let.* (PDT)

The downward trend in the birth rate is not a recent phenomenon, ***but***$_{\textbf{correction}}$ *one that has significantly worsened since the mid-60s.*

(131) *Improvizace je dobrá věc, **ale**$_{\textbf{restr. opposition}}$ je potřebné se zamyslet nad možnými eventualitami a důsledky.* (PDT)

*Improvisation is a good thing, **but**$_{\textbf{restr. opposition}}$ it is necessary to consider the possible eventualities and consequences.*

Contrary to our assumptions that syntactic *Adversative* relation corresponds to discourse relations from the CONTRAST class only, there is also 1% of cases, where some other discourse types were annotated for structures with *Adversative* relation. Namely, *conjunction*, *asynchrony*, *synchrony* and *explication* have been found. Examples of these structures are discussed in the next subsection, since they represent cases where syntactic and discourse-level analyses differ in their perspective (cf. also Example 135).

Syntactic analysis captures semantic relations between clauses from a different perspective than discourse-level analysis

Contrary to all structures discussed so far, all other cases where there is a mismatch between syntactic and discourse-level analyses cannot be generalized. These cases present contexts where annotators decided to annotate some other discourse relation than the corresponding one present in syntactic analysis. Such cases clearly illustrate that syntactic and discourse-level perspective are different types of linguistic analysis. Because of their unsystematic character, these cases cannot be easily enumerated. Thus, we only illustrate them using Tables 9.2 and 9.3 and Examples 132–135.

Table 9.2 displays all types of discourse relations corresponding to the syntactic *Conjunction* (labelled *CONJ*). The first row in the table represents the most frequent and regular case where the tectogrammatical label *CONJ* corresponds to discourse *conjunction*,[94] the rest of the table illustrates cases where annotators decided to mark discourse relations other than *conjunction*.

The contexts where structures with syntactic *Conjunction* were newly annotated from the discourse-level perspective are provided in Examples 132–134. They illustrate discourse relations of *reason–result*, *instantiation* and *generalization*, respectively, and thus clearly indicate the semantic orientation of discourse-level analysis in contrast to a more form-based syntactic treatment of these structures.

(132) *Tetování po několika letech bledne **a**$_{\textbf{reason–result}}$ je třeba jej obnovit.* (PDT)

*Tattoos fade after a few years **and**$_{\textbf{reason–result}}$ must be renewed.*

(133) *Čas tráví učením a prací, **například**$_{\textbf{instantiation}}$ opravují objekty, ve kterých jsou ubytováni.* (PDT)

*They spend their time studying and working, **for example**$_{\textbf{instantiation}}$ they repair buildings in which they are staying.*

[94] Similarly, in the majority of cases of correspondence between the syntactic label *ADVS* for *Adversative* relation and discourse-level *opposition*, these cases were also left aside for automatic extraction of discourse relations (see the next subsection).

Type of discourse relation	Number of occurrences
conjunction	5,198
specification	208
asynchrony	159
reason–result	140
correction	86
gradation	64
confrontation	60
opposition	60
explication	39
conjunctive alternative	26
synchrony	24
equivalence	18
disjunctive alternative	16
concession	16
instantiation	14
restrictive opposition	13
generalization	8
condition	7
pragmatic reason–result	5
pragmatic condition	1

Table 9.2: Discourse relations corresponding to structures with the syntactic *Conjunction* (*CONJ*) on the tectogrammatical layer

(134) *V Africe lidé umírají hladem, všeobecně chátrá životní prostředí, někdo se musí postarat o bezprizorní děti –* ***prostě***$_{\textbf{generalization}}$ *je strašně moc věcí, které tyto peníze potřebují.* (PDT)

In Africa, people are dying of hunger, the environment is generally deteriorating, someone has to take care of the street children – there is ***just***$_{\textbf{generalization}}$ *an awful lot of things that need that money.*

A similar situation can also be found in other syntactic structures, e.g. in structures with syntactic *Adversative* relation (*ADVS*) that have been already mentioned. Example 135 illustrates the discourse relation of *conjunction* in construction which is presented as syntactically *Adversative*.

(135) *Redaktor Adámek nebyl včera k dosažení, jeho rodinní příslušníci* ***však***$_{\textbf{conjunction}}$ *Zemanem uváděné souvislosti jeho odchodu z televize také nepotvrdili.* (PDT)

Editor Adámek could not be reached yesterday, ***however***$_{\textbf{conjunction}}$*, his family members also did not confirm the circumstances of his departure from television, as they were presented by Zeman.*

Syntactic labels for coordinate structures	Discourse relevant occurrences	Different discourse interpretation (%)
Conjunction (*CONJ*)	6,159	15
Gradation (*GRAD*)	131	14
Reason (*REAS*)	214	6
Consequence (*CSQ*)	342	4
Disjunction (*DISJ*)	259	3
Adversative relation (*ADVS*)	1,479	1
Confrontation (*CONFR*)	14	0

Table 9.3: Semantically different discourse interpretation of syntactic labels for coordination

To provide illustration for the frequency of cases where syntactic analysis differs from the discourse-level one in our data, we measured (in percent) how frequently these cases occur for all coordinate labels.[95] The summary is given in Table 9.3. The second column displays the number of occurrences for all coordinate labels between clauses in the PDT, the third column represents the percentages of cases where these syntactic labels do not correspond to discourse labels.[96]

As these results indicate, structures with syntactic *Conjunction* and *Gradation* (syntactic labels *CONJ* and *GRAD*) were most frequently interpreted as some other relations in the discourse-level analysis. On the other hand, structures with syntactic *Adversative* relation and *Confrontation* (syntactic labels *ADVS* and *CONFR*) were almost never newly annotated. While this result could be expected for syntactic *Conjunction* due to semantic vagueness of this notion, we are not sure what causes the relatively big mismatch between *Gradation* at syntactic layer and its discourse counterparts. It may be the result of a form-oriented decision in syntactic annotation which was conducted by the main lexical meaning of conjunctions. Syntactic *Gradation* has distinctive connectives in Czech (e.g. *nejen – ale i* [*not only – but also*], *dokonce* [*even*], *navíc* [*moreover*]) and it may be the case that they do not convey discourse relation of *gradation*. However, overall, we can see that the syntactic labels for coordination correspond to their discourse counterparts significantly (from 85 to 100% of cases).

[95] These are labels used for representing a connection of coordinate clauses in a tree, cf. e.g. the node with the label *CONJ* in Figure 9.1. All discourse relevant coordinate syntactic labels are listed in the first column of Table 9.3.

[96] For the *Adversative* label (*ADVS*), all cases presented in Table 9.1 are treated here as corresponding, since they only represent a finer classification of the *ADVS* relation not a semantically different interpretation.

Syntactic label	Type of discourse relation
Subordinate structure	
Aim (*AIM*)	*purpose*
Cause (*CAUS*)	*reason–result*
Concession (*CNCS*)	*concession*
Condition (*COND*)	*condition*
Contradiction (*CONTRD*)	*confrontation*
Substitution (*SUBS*)	*correction*
Coordinate structure	
Adversative relation (*ADVS*)	*opposition*
Confrontation (*CONFR*)	*confrontation*
Conjunction (*CONJ*)	*conjunction*
Consequence (*CSQ*)	*reason–result*
Disjunction (*DISJ*)	*disjunctive alternative*
Gradation (*GRAD*)	*gradation*
Reason (*REAS*)	*reason–result*

Table 9.4: Syntactico-semantic label to discourse type automatic translation table

Syntactic analysis directly transferable to discourse-level analysis

So far, we have been describing cases where the syntactic analysis somehow differs from discourse-level analysis. This subsection on the other hand discusses cases where tectogrammatical labels correspond directly to their discourse-level counterparts and thus could be exploited during automatic extraction of discourse relations. Table 9.4 shows how discourse-relevant syntactic labels were converted into discourse labels.

The first six rows represent subordinate relations (i.e. relations between dependent and their governing clauses), the last seven rows represent coordinate relations (i.e. relations between coordinated clauses). It should be noted that in some cases there is no one to one correspondence between the syntactic label and its discourse counterpart. For example, the discourse relation *reason–result* corresponds to three relevant syntactic labels: cause (*CAUS*, for subordinated causal clause), consequence (*CSQ*, for coordinated structure where result or consequence is described in the second clause) and reason (*REAS*, for coordinated structure with cause or reason expressed in the second clause). In the discourse analysis in the PDT, these syntactic relations were merged into one discourse relation, because the placement of argument where reason versus result is expressed is indicated by the direction of the discourse relation (i.e. the arguments are distinguished semantically, cf. Chapter 2).

Type of the relation	Number of occurrences
intra-sentential relations	12,623
automatic annotation	9,663
semi-automatic annotation	491
manual annotation	2,469
inter-sentential (all manual)	5,538
total	18,161

Table 9.5: Overview of discourse relations annotated in the PDT

To show the utilization of syntactic labels for discourse-level analysis, we can compare overall numbers of relations annotated in both phases of annotation: manual and automatic. This summary is displayed in Table 9.5. The distinction manual versus automatic annotation is only valid for intra-sentential discourse relations (i.e. those occurring within a single sentence), because the syntactic analysis in the PDT in principle does not surpass sentence boundaries and thus all inter-sentential discourse relations (i.e. relations between separate sentences) were annotated manually. Because of a rich variety of connectives, some manual work preceded in the case of temporal relations (491 relations), these cases are treated as semi-automatic in Table 9.5 and are not counted into fully automatic category. The total number of all discourse relations is also presented.

The overview given in Table 9.5 indicates that 53% of all discourse relations (i.e. 9,663 of 18,161 relations) were extracted automatically using features from syntactic analysis on the tectogrammatical layer of the PDT. The percentage is even higher when we take into account only intra-sentential relations (i.e. those which are fully covered by syntactic analysis) almost 80% of discourse relations expressed within a single sentence (i.e. 9,663 of 12,623 relations) were extracted automatically. However, as we already mentioned, the exact contexts where syntactic labels also convey discourse meaning were identified manually by annotators (they left these places aside for the automatic phase of annotation).

These results illustrate that there is still a difference between syntactic and discourse-level analysis: One fifth of intra-sentential relations were treated differently from the syntactic versus discourse points of view. This information is quite interesting, especially for its theoretical consequences – it was shown using the same data that syntactic and discourse analysis are different tasks even if we consider only the relations between clauses in a single compound sentence.

9.1.2 Scope of discourse arguments

Apart from the syntactico-semantic labels for relations between clauses (discussed so far in Section 9.1.1) other features of syntactic analysis on the tectogrammatical layer also turned out to be useful during both phases of discourse annotation. In the remaining part of Section 9.1 we discuss these features in detail proceeding from the scope of argument via resolution of coordination and ellipsis to the identification of connectives.

Effective subtree as a discourse argument

In approaches where annotation was carried out on plain texts (i.e. on not syntactically annotated material), the scope of argument is defined as a text span containing the amount of information that is minimally required and at the same time sufficient to complete the semantics of the relation (cf. e.g. Prasad et al., 2008). In our approach, with discourse relations annotated on top of dependency trees, the scope of the arguments is defined as a set of nodes representing an amount of language material sufficient for the interpretation of a discourse relation. However, it turned out during the manual phase of annotation, that the scope of an argument in a vast majority of cases corresponds to the *effective subtree* of the root node of the argument (i.e. set of nodes that linguistically depend (transitively) on the given node, taking all effects of coordinations etc. into account). The root node of the argument can either be a finite verb or a node coordinating finite verbs. At the same time, all nodes belonging to the other argument of the relation are excluded from this subtree. Figure 9.2 shows this principle on the *conditional* relation represented in Figure 9.1 above (p. 119). One argument is represented by a subtree under the node with label *CONJ* connecting verbs *to want* and *to be born*, the second argument is then created by a subtree under the node with label *CONJ* coordinating verbs *to develop* and *to be* but excluding the subtree of the first argument. This definition of scope turned out to be useful in automatic annotation as well.

A manual random check of automatically annotated relations proved that in automatically (and semi-automatically) annotated intra-sentential relations, the tectogrammatical tree structure correctly defined the scope of the arguments, independently of whether the argument was formed by a continuous sequence of words on the surface level, or not.

For all manually annotated relations (2,469 intra-sentential cases and 5,538 inter-sentential, cf. Table 9.5) in all but 146 cases the scope of arguments is also equal to the effective subtree of the root node. Thus, we can claim, that the tree structure captures the scope of discourse argument reliably (in 99% of cases). In the 146 cases, the annotator defined a different scope of the argument, for example since some clauses in complicated structures revealed to be redundant for the interpretation of the discourse relation. From the beginning, a special attribute was established

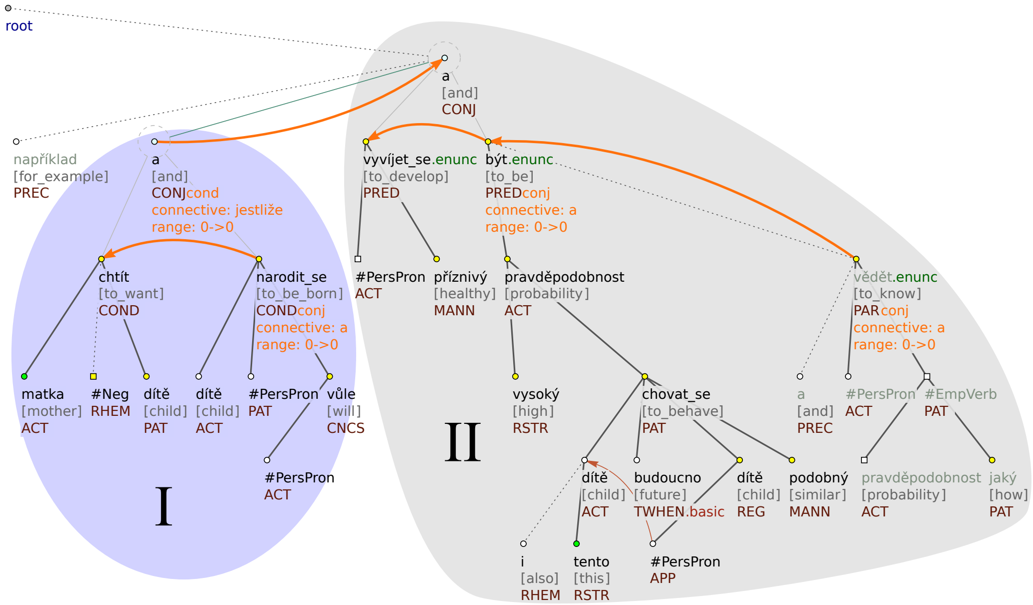


Figure 9.2: The tectogrammatical representation of Example 118 (p. 118) with marked arguments of a discourse relation of *condition*

for this situation which allows the annotator to join any set of nodes into a group. An illustration of this situation is given in Example 136. These two compound sentences are connected through a discourse relation of *opposition* anchored by *však* [*however*]. Nevertheless, the underlined material is not necessary for discourse relation interpretation and therefore it was excluded from the arguments by means of joining all other nodes in these structures into groups and connecting these groups by discourse relation.

(136) *Přestože klusácký sport u nás dosahuje při srovnání zahraničních startů daleko větší úspěšnost než cvalový provoz,* <Arg1: *stál pro téměř neexistující propagaci na okraji zájmů dostihové veřejnosti.*> *Jak* ***však***$_{\textit{opposition}}$ *včera zaznělo na tiskové konferenci,* <Arg2: *měla by se tato situace letos změnit k lepšímu, protože ČKA rovněž podepsala smlouvu o marketingu se společností Impact (dceřiná společnost Art production K.*> (PDT)

Although harness racers in our country are far more successful than gallop racers when compared to foreign events, <Arg1: *it was of marginal interest to the racing public, given the almost non-existing promotion.*> ***However***$_{\textit{opposition}}$*, as was said yesterday at a press conference,* <Arg2: *this situation should change for the better this year because the ČKA also signed a contract with the company Impact (subsidiary of Art production K.) for marketing services.*>

Coordination resolution

Together with a general tree structure, two other tectogrammatical features helped both manual and automatic annotation – namely coordination and ellipsis resolution.

In the PDT, coordinating expressions are represented as separate nodes (as seen in Figure 9.1 – there are two nodes for coordination, both with syntactic label *CONJ* for syntactic *Conjunction*). These nodes were exploited both in manual and automatic annotation. In the automatic detection of discourse arguments, the procedure always searched for the topmost suitable coordination. For example, two coordinated conditional clauses in Figure 9.1 (the subtree under the node *to_want* and *to_be_born*) and Example 118 create together a single complex discourse argument. Therefore, instead of two discourse relations that could apply directly between the individual verbal nodes (i.e. between each conditional clause and the second argument), only one overall discourse relation was established using the node with label *CONJ* coordinating both conditional clauses as a starting point for the argument representation. This is a more comprehensible solution, without any loss of information.

Another illustration is given in Example 137 and in Figure 9.3 – the *disjunctive alternative* relation (discourse label *disjalt*) holds between the clauses *he could join, for example, Charter 77*, *contribute to November 1989* and the clause *he could do what he considered to be more beneficial for him*. Instead of two discourse relations that could be created directly between the individual subtrees under verbal nodes (i.e. between the subtrees under the nodes *to_join* and *to_do*, between the subtrees under the nodes *to_have_merit* and *to_do*), only one overall discourse relation was created by the automatic procedure (between the subtree under the node for coordination with label *CONJ* and the subtree under the verb *to_do*). The topmost suitable coordination – in this case the node with syntactic label *CONJ* (used for *Conjunction* in syntactic analysis) – was searched for.

Similarly, for the *specification* relation (the topmost relation in the tree in Figure 9.3), the topmost coordination (with syntactic label *DISJ* used for *disjunctive* relation) was used instead of two lower nodes (the node *and* and the node *to_do*). In this case, the type of discourse relation was assigned manually – syntactic *Conjunction* was interpreted here as *specification* in discourse-level analysis.

(137) <Arg1: *Mohl si v té době vybrat*> –$_{\text{specification}}$ <Arg2: <Arg1: <Arg1: *připojit se například k Chartě 77*> ***a***$_{\text{conjunction}}$ <Arg2: *zasloužit se o listopad 1989,*>> ***nebo***$_{\text{disjunctive alternative}}$ <Arg2: *udělat to, co pokládal za výhodnější.*>> (PDT)

<Arg1: *He could choose at that time*> –$_{\text{specification}}$ <Arg2: <Arg1: <Arg1: *he could join, for example, Charter 77*> ***and***$_{\text{conjunction}}$ <Arg2: *contribute to November 1989,*>> ***or***$_{\text{disjunctive alternative}}$ <Arg2: *he could do what he considered to be more beneficial for him.*>>

There were almost 1,000 cases where this shift to the topmost suitable coordination made automatic annotation more comprehensible in our data. From the point of view

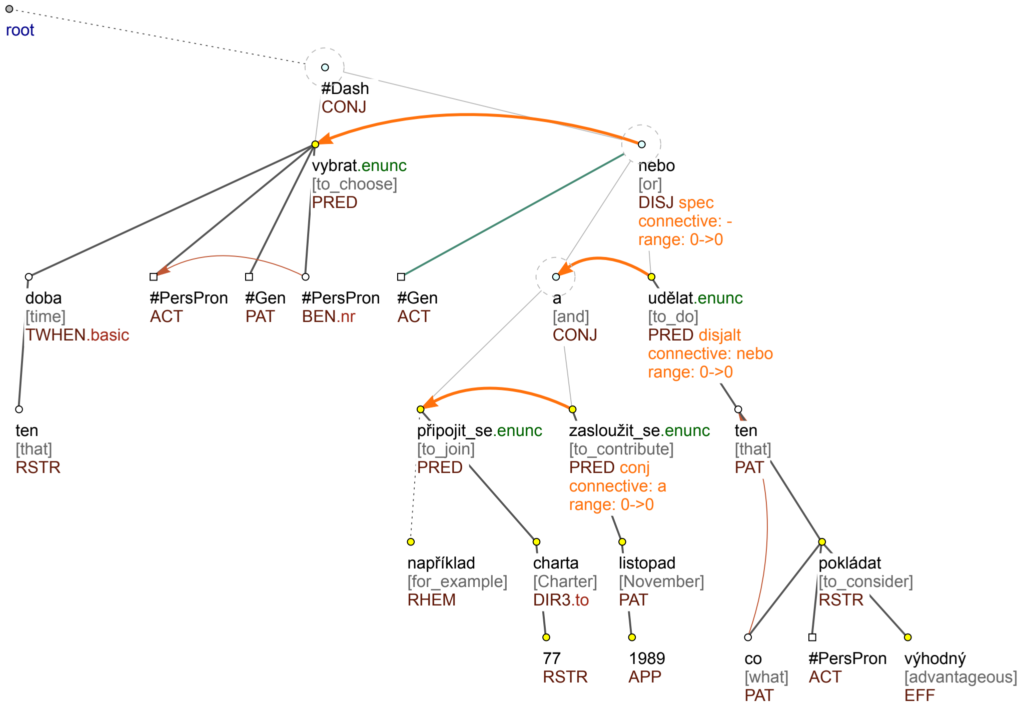


Figure 9.3: A tree with complex coordination, representation of Example 137

of manual annotation, the tree structure and the coordination resolution especially help annotators to decide about the scope of arguments, particularly in syntactically complicated cases.

Ellipsis resolution

In structures with an ellipsis in the surface form of the sentence, missing expressions have been reconstructed on the tectogrammatical layer of the PDT.[97] This reconstruction was helpful in both phases of discourse annotation (manual as well as automatic), namely in case of reconstructed verbal nodes. Thus, we were able to mark 1,630 relations that have an elided verb in one or both arguments. Without the ellipsis resolved, the relations could be easily overlooked in the text or it would not be possible to annotate them in the trees at all. Example 138 shows discourse relation of *correction* in structure with an elided verb. The structure contains two clauses on the

[97] In the PDT, deleted nodes are reconstructed in the case of grammatical incompleteness of the surface form only (for detailed analysis see Mikulová, 2011, for some examples of reconstructed nodes see Chapter 3, Section 3.5.3).

tectogrammatical layer – the finite verb in the second of them is a reconstructed verb *to go*.

(138) *Zloději nechodí po horách, **ale**$_{\textbf{correction}}$ [chodí] po domácnostech.* (PDT)
*Thieves do not go to mountains **but**$_{\textbf{correction}}$ [they go] to houses.*

9.1.3 Connectives

As we had expected before the annotation, the tectogrammatical layer provided several useful clues for connective identification in both manual and automatic phase of annotation. In this section, we speak about the manual phase in the first instance, then features used during the automatic phase are discussed.

Expressions referring to preceding context (syntactic label *PREC*)

In the manual phase of annotation, annotators went through plain (printed) texts, marked an open set of possible connectives and then viewed trees and annotated discourse relations and the scope of discourse arguments where suitable. In this phase, expressions marked with the syntactic label *PREC* (a reference to PREceding Context) appeared to be a useful clue for a possible presence of a discourse relation. In the syntactic annotation, this syntactic label was adopted for expressions that signal (typically) an inter-sentential relation during annotation of the tectogrammatical layer (which otherwise does not in principle surpass sentence boundaries). Such expressions are e.g. *proto* [*therefore*], *ovšem* [*however*], *tedy* [*hence*]. Nevertheless, this label neither interprets the semantic type of relation, nor specifies the scope and the position of the other discourse argument. An example of an expression with this label (i.e. *například* [*for example*]) can be seen in Figure 9.1 above (p. 119).[98]

Expressions with label *PREC* were drawn on not only during discourse annotation but also after each part of the annotation to check its completeness. The total number of occurrences of expressions with the syntactic label *PREC* in our data is approximately 5,500. The vast majority (approximately 4,300) were annotated as discourse connectives (3,900 in inter-sentential relations, 400 in intra-sentential relations). A sentence with an intra-sentential expression referring to preceding context is given in Example 139 – expression *pak* [*then*] is a part of the connective *jestliže – pak* [*if – then*].

(139) ***Jestliže** cestující musí náhradou za přerušené dopravní spojení použít více dopravních prostředků, **pak**$_{\text{PREC}}$ stále platí jízdenka z prvního prostředku i na ostatních linkách.* (PDT)

***If** a passenger has to use several means of transport as a detour for a disrupted transport link, **then**$_{\text{PREC}}$ the ticket from the first mode of transport is still valid for the others.*

[98] As was already mentioned, this sentence explains the meaning of the notion of pathology of the third generation that appeared in a previous context.

The remaining occurrences of expressions syntactically labeled as *PREC* were marked by an annotator's comment in the data; in most cases they are signals of writer's attitude to the content of the text.

Rhematizing particles (syntactic label *RHEM*)

There is also another group of expressions, namely rhematizing or focusing particles like *také* [*too, also*], *i* [*also*] or *jenom* [*only*], which participate in establishing discourse coherence.[99] Originally, we assumed that these expressions, marked with syntactic label RHEM, would be used in the search for discourse connectives as well. However, contrary to our expectation, it was found out that from the discourse point of view, these expressions are the most ambiguous elements. Although we established criteria for distinguishing rhematizers in a connective function from other uses, in some contexts the difference is not clear enough. The main principle adopted for interpreting a rhematizers also as a discourse connective is as follows: rhematizers function as discourse connectives if they connect two text spans containing finite verbs with different meanings. The first condition (connecting two text spans) is valid for all connectives, the second one (different meaning of verbs) was added after the observation that rhematizers very often connect only nominal groups, rather than full discourse arguments. Cf. Example 140, where the expression *také* [*also*] connects the nominal group *Chile* with the nominal group *partners from Argentina*.

(140) *Letos by měli argentinští partneři odebrat asi 12 tisíc motocyklů za celkem 12 mil. dolarů. Předběžné kontrakty na dodávky motocyklů a jízdních kol získal podnik* ***také***$_{\text{RHEM}}$ *z Chile.* (PDT)

This year, the Argentinian partners should receive about 12,000 motorcycles for a total of 12 mil. dollars. The company obtained some preliminary contracts for the supply of motorcycles and bicycles ***also***$_{\text{RHEM}}$ *from Chile.*

To be sure that a rhematizer connects verbal groups, not mere nominal groups, different meanings of the verbs in connected text spans is required.[100] According to this principle, the expression *také* [*also*] in Example 141 is considered to be a discourse connective (the meanings of predicates *podporovat* [*to support*] and *být spokojen* [*to be satisfied*] are different), while in Example 142 the word *také* [*also*] is not considered to be a discourse connective, since the predicates *zůstat zavřeno* [*to remain closed*] and *neotevřít* [*to be not opened*] are semantically the same (and the rhematizer thus connects mere nominal groups). In cases with similar meanings of predicate verbs, the synonymous verb in the second text span is to a certain degree semantically redundant and both text spans can be thus reformulated as containing only one verb (cf. reformulation of

[99] For a comprehensive treatment on these expressions see Štěpánková (2014), for discussion on relation between rhematizers and discourse connectives see Mladová (2008).

[100] However, we are aware of the fact that rhematizers connecting nominal groups contribute significantly to the coherence of the discourse as well.

Example 142: *Around 70 percent of stands and also most of the shops remained closed in the morning.*). The rhematizer is therefore considered to connect mere nominal groups.

(141) *Dánsko plně podporuje budoucí zapojení ČR do Unie. Obě strany jsou* ***také***$_{\text{RHEM}}$ *spokojeny s růstem vzájemného obchodu.* (PDT)

Denmark fully supports future integration of the Czech Republic into the European Union. Both sides are ***also***$_{\text{RHEM}}$ *satisfied with the growth of bilateral trade.*

(142) *Kolem 70 procent stánků na tržištích zůstalo zavřeno a* ***také***$_{\text{RHEM}}$ *většina obchodů dopoledne neotevřela.* (PDT)

Around 70 percent of the stands at the market remained closed and ***also***$_{\text{RHEM}}$ *most of the shops did not open in the morning.*

As these two examples indicate, the principle discussed above was adopted for intra-sentential uses (within a single sentence) of rhematizers as well as for inter-sentential ones (between sentences).

Besides functioning as discourse connectives, some rhematizers also display other types of meaning – for example, they express various attitudes of the speaker as in Example 143. In our approach, these uses are not relevant for annotation of discourse relations although they are very interesting for other types of text analysis.

(143) *Byli jsme strašně nervózní a podle toho naše hra* ***také***$_{\text{RHEM}}$ *vypadala.* (PDT)

lit. *We were very nervous and according to it our game* ***also***$_{\text{RHEM}}$ *looked like.*

We were very nervous and we ***really*** *did play accordingly.*

Automatic detection of discourse connectives

In the automatic phase of annotation, the discourse connectives of intra-sentential discourse relations could be automatically detected on the basis of the information from the syntactic analysis. Connectives of subordinate relations (e.g. *když* [*when*], *protože* [*because*]) could be found among different nodes from the surface syntactic layer (a-layer, see Chapter 6) that correspond to the verbal root of the discourse argument on the tectogrammatical layer. Connectives of coordinate relations could be found on the tectogrammatical layer at the coordinate node (see e.g. connective *a* [*and*] in the top-most coordinate node in Figure 9.1) or its modifiers (expressions as *dokonce* [*even*] or negation are treated as modifiers).

With the exception of 23 atypical cases (which were fixed manually), discourse connectives could be detected automatically for all 10,154 intra-sentential discourse relations left aside for the automatic phase of annotation (using information already present in syntactic analysis). Thus we can claim that for connective identification of intra-sentential relations, syntactic analysis was really helpful.

However, as was already mentioned in connection with automatic extraction of discourse types using syntactic labels, in contexts where syntactic and discourse-level analyses did not correspond, the connectives were detected manually.

Expression *což* [*which*]

As a supplement of this discussion, let us point out one interesting case of the connective detection. The expression *což* (roughly [*which*]) with pronominal origin[101] represents an intra-sentential connective with the meaning of *conjunction* and is, at the same time, inflected and plays a role of a participant in the clause structure. To distinguish the connective role of this expression automatically, the feature of grammatical coreference[102] was used. The deictic part of the expression *což* can refer to a verbal group (e.g. *Pavlov then became the prime minister* in Example 144) or to a nominal group (*a love for war* in Example 145). However, it has the role of a connective only in the first case, i.e. when it refers to a verbal group as a full discourse argument (Example 144).

(144) *Pavlov se pak* ***stal předsedou vlády****,* ***což*****$_{\text{conjunction}}$** *se Klausovi přihodilo nakonec také.* (PDT)

Pavlov then ***became the prime minister****,* ***which*****$_{\text{conjunction}}$** *after all happened to Klaus as well.*

(145) *Cítil jsem z nich* ***lásku k válce****,* ***což*** *je něco proti přírodě.* (PDT)

I felt ***a love for war from them****,* ***which*** *goes against nature.*

There is a total of 355 occurrences of the expression *což* in our data, 220 occurrences have a grammatical coreference link to a finite-verb node, 11 occurrences have this link to a coordination of finite-verb nodes. Therefore, thanks to the grammatical coreference, it was possible to automatically distinguish these 231 (220+11) occurrences from the rest and identify the expression *což* as a discourse connective in these contexts.

9.2 Discourse Structure from the Syntactic Point of View

While we discussed the features of syntactic analysis which provide help or point to the need of new annotation in the process of marking discourse relations in Section 9.1, in Section 9.2, we look at the relation of syntactic analysis and discourse-level analysis from the opposite point of view – we want to see what insights can

[101] It has arisen from the relative pronoun *co* [*what*], which is inflected according to its role in a sentence, and the bound particle *-ž*.

[102] Grammatical coreference has been annotated in the PDT for expressions for which it is possible to identify the coreferred part of the text on the basis of grammatical rules. This applies e.g. for relative pronouns, reflexive pronouns or for participants of control verbs (see Chapter 3 and also Mikulová et al., 2006).

syntactic analysis reveal about discourse relations. This section focuses on two main topics: first, the inter-relation between syntactic structure and discourse structure is defined with a consideration for the distribution of discourse relations realized within a single sentence versus between sentences (9.2.1), then we take a closer look at those discourse relations realized within a single sentence and describe their distribution in subordinate or coordinate structures (9.2.2).

9.2.1 Discourse relations realized intra-sententially and inter-sententially

Let us begin our reflections on syntactic characteristics of discourse relations by two observations. The first of them is based on Table 9.5 (p. 131), which displays an overall characteristics of discourse relations according to their realization either in one compound sentence (intra-sentential relations) or between sentences (inter-sentential relations). We see that in our data there are 12,623 intra-sentential relations, while inter-sentential relations are much less frequent – there are only 5,538 of them. Syntactic analysis thus reveals that when expressed by discourse connectives (and our annotation concerns only these cases), discourse relations are more frequently realized in one compound sentence than between separate sentences.

The second observation concerning the relation between syntactic and discourse analyses takes into account each discourse semantic type separately. Table 9.6 displays the distribution of all semantic types in intra- versus inter-sentential realization. The second column of the table shows the total number of occurrences for each semantic type in the PDT, the rest of the table displays percentages of intra-sentential and inter-sentential realizations. Semantic types are ranked according to the percentage of realization within a single sentence.

According to our understanding of discourse relations, there is a necessary condition for any relation between clauses to be considered a discourse relation: it must be possible to express this relation between two independent text spans (i.e. between two separate sentences, not only intra-sententially). From this perspective, all semantic discourse types are the same – it is possible to relate two independent text spans through these relations. However, in the course of data development we noticed that different discourse semantic types, in fact, act differently in this respect.

Thus, we formulated a hypothesis that there may be a certain scale that determines to what extent each relation is preferably expressed either within a single sentence or between sentences. Further, a question arose whether relations are grouped in a certain way according to this property. After processing the whole data set, we observed that the first expectation was confirmed – the numbers represent a scale with continuous transition rather than some distinct groups of semantic types (as Table 9.6 indicates). We found more semantic discourse types which are preferably realized within a single sentence (cf. lines from *purpose* to *reason–result*) than types preferably realized between sentences (cf. lines from *opposition* to *generalization*) – the scale is thus rather non-symmetric. We can also see that while there are four types

Type of discourse relation	Number of occurrences	Intra-sentential (%)	Inter-sentential (%)
purpose	373	99	1
condition	1,185	98	2
disjunctive alternative	246	95	5
pragmatic condition	15	93	7
specification	554	81	19
conjunction	6,619	81	19
conjunctive alternative	78	79	21
synchrony	186	76	24
correction	409	73	27
concession	794	70	30
asynchrony	723	70	30
reason–result	2,325	61	39
confrontation	584	53	47
gradation	383	51	49
pragmatic opposition	48	45	55
opposition	2,828	43	57
explication	213	43	57
equivalence	95	40	60
restrictive opposition	236	36	64
pragmatic reason–result	39	30	70
instantiation	134	19	81
generalization	92	8	92
total	18,161	70	30

Table 9.6: Overview of discourse relations occurring within a single sentence or between sentences in the PDT

which are realized in more than 90% within a single sentence, there is only one which is realized between sentences in the same percentage of occurrences (*generalization*).

Regular syntactic counterparts of discourse semantic types

If we look at possible regular syntactic counterparts for each discourse type, we see that all relations with a regular subordinate form in Czech (i.e. *condition*, *purpose*, *synchrony*, *asynchrony*, *concession*, *reason–result*) occupy places in the top half of Table 9.6 or, in other words, they are more often expressed within a single sentence than between sentences, while relations with a regular coordinate form in Czech (*conjunction*, *correction*, *confrontation*, *opposition*, *gradation*, *conjunctive* and *disjunctive alternative*) do not form such a group. However, at the same time, relations with a regular subordinate form distinctly differ from each other in the proportion of intra-sentential and inter-sentential forms (cf. for example the relations of *condition* and of *concession*).

Comparison of semantic types from the same semantic class

Table 9.6 further allows for a comparison of semantic types belonging to the same semantic class of discourse relations (see Chapter 2). Both TEMPORAL relations (*synchrony* and *asynchrony*) act very similarly (according to our expectation) and also CONTRAST relations (*correction*, *concession*, *confrontation*, *gradation*, *pragmatic opposition* and *opposition*) form a rather continuous group in the table.

On the other hand, relations from the CONTINGENCY class (*purpose*, *condition*, *pragmatic condition*, *reason–result*, *explication* and *pragmatic reason–result*) are scattered throughout the whole table.

The last semantic class, EXPANSION, is distributed between two different parts. The first one (*disjunctive alternative*, *specification*, *conjunction* and *conjunctive alternative*) is placed in the top half of the table, whereas the second part (*equivalence*, *instantiation* and *generalization*) appears at the very bottom of the table. If we take a closer look at the EXPANSION class, it is apparent that from the informational viewpoint, this class consists of relations which continue discourse by adding new information (*conjunction* and both *alternatives*) and of relations continuing discourse by expressing something that was already said in a different way (*specification*, *equivalence*, *instantiation* and *generalization*). The second group was expected to occur mostly between separate sentences; and with *specification* forming an exception, this is true for the presented data. In our opinion, the unexpected placement of *specification* among semantic types which are preferably realized within a single sentence results from the fact that only relations anchored by explicit connectives are taken into account in our annotation. The PDTB data seem to confirm that *specification* is very often realized without a connective (Prasad et al., 2007, pp. 75 and 90). Thus, we may assume that when all relations (not only explicit ones) are taken into consideration, *specification* would be placed near *equivalence*, *instantiation* and *generalization*.

Discourse relations *conjunction* and *opposition*

From a purely linguistic viewpoint, the distinction between *conjunction* and *opposition* in intra- versus inter-sentential realizations is very interesting. We would expect that these two types would act similarly, but the data revealed that relation of *conjunction* is realized primarily within a single sentence, while *opposition* shows a balanced distribution of intra- and inter-sentential realizations. We can assume that this difference follows from the fact that *opposition* needs to be signalled with a connective when realized between sentences while *conjunction* can be easily expressed by merely placing sentences one after another (without any connective).

9.2.2 Discourse relations in subordinate versus coordinate structures

After having discussed basic distribution for each semantic type of discourse relations within a single sentence and between sentences we can now turn to an area where discourse-level analysis meets syntactic analysis most frequently and most closely. This section focuses only on intra-sentential relations and discusses their distribution in subordinate versus coordinate structures.

We begin the discussion with two examples of the *asynchrony* relation in order to clarify what is meant by the terms of subordination and coordination. In Example 146, the *asynchrony* relation is realized in a subordinate structure, while Example 147 displays it in a coordinate construction.

(146) *Neznámý pachatel pustil bankovky do oběhu v místní obchodní síti,* ***když*** *na ně* ***předtím***$_{\textbf{asynchrony, subordinate structure}}$ *nalepil kolky ze stokorun.* (PDT)

An unknown offender released banknotes into circulation in the local shopping network, ***when*** *he* ***first***$_{\textbf{asynchrony, subordinate structure}}$ *had stuck stamps from one hundred banknotes on them.*

(147) *Výstava bude otevřena do 20. října,* ***pak***$_{\textbf{asynchrony, coordinate structure}}$ *poputuje do Bratislavy a dalších evropských měst.* (PDT)

The exhibition will be open until October 20, ***then***$_{\textbf{asynchrony, coordinate structure}}$ *it will go to Bratislava and other European cities.*

The whole distribution of all the intra-sentential discourse relations in subordinate and coordinate constructions is illustrated by Table 9.7.[103]

As all semantic types are distinguished in Table 9.7, we can observe if and how some types differ from others and subsequently we can offer some linguistic explanations.

Prevalence of coordinate structures

Based on the presented results, the following syntactic characteristics of intra-sententially realized discourse relations can be stated: first, when realized within a single sentence, discourse relations are related rather to coordinate than to subordinate structures (coordinate structures cover 70% of all intra-sentential realizations of discourse relations). This feature arises as a consequence of the nature of discourse relations – they are connected with the text structure rather than with the sentence structure. A connection of text spans in a coordinate compound sentence is more similar to a connection of text spans across sentence boundary than it is the case for the connection

[103] The total number of intra-sentential discourse relations does not correspond to Table 9.5, since structures containing a parenthesis (there is 87 of them in our data) as a discourse argument were excluded from the analysis. Parenthesis is considered to be a phenomenon outside the coordinate versus subordinate distinction.

Type of discourse relation	Intra-sentential	In coordinate structures (%)	In subordinate structures (%)
conjunctive alternative	62	100	0
generalization	8	100	0
disjunctive alternative	234	99	1
conjunction	5,329	99	1
gradation	197	96	4
opposition	1,221	95	5
correction	298	93	7
equivalence	37	92	8
instantiation	22	82	18
specification	446	75	25
restrictive opposition	84	76	24
explication	87	68	32
pragmatic reason–result	12	67	33
pragmatic opposition	22	64	36
reason–result	1,415	48	52
confrontation	312	44	56
asynchrony	503	34	66
synchrony	143	18	82
concession	554	18	82
pragmatic condition	12	8	92
condition	1,165	1	99
purpose	373	0	100
total	12,536	70	30

Table 9.7: Overview of intra-sentential discourse relations occurring in coordinate and subordinate structures in the PDT

of text spans in subordinate structure. In other words, this proportion in our data correspond to our assumption that subordinate structures are a domain for expressing complements of a verb (including obligatory ones) in a governing clause rather than a domain for building text structure.

On the other hand, the proportion of discourse relations realized in subordinate constructions shows that both domains (the domain of building text structure and domain of sentence structure represented by subordinate constructions) partially overlap. We can see that this overlap is particularly large for discourse relations of *purpose*, *condition*, *pragmatic condition*, *concession* and *synchrony*, which are realized in more than 80% of cases in subordinate structures – in other words, these relations tend to be the most syntax-bound.

Discourse relations with semantically mixed nature

Secondly, these results can contribute to the discussion concerning the nature of some relations. From the theoretical point of view adopted in our approach, especially two relations seem to have a rather mixed nature, i.e. they embody the semantic characteristics of two general classes of discourse relations simultaneously. These relations are *concession* and *explication*. *Concession* is often classified as one of the CONTRAST relations because there is a strong contrast expressed between its arguments, but at the same time, it can be seen as a CONTINGENCY relation due to the violation of an expected causal connection between the contents of the arguments. Consider e.g. Example 148. Based on the fact that a new football season is in full swing, we normally expect that affairs of the last years are completely over. However, this expectation is violated here – players have not received their money yet.

(148) ***Ačkoliv***$_{\textbf{concession}}$ *je fotbalová liga v plném proudu, hráči Sparty dosud nedostali prémie za mistrovský titul v loňském ročníku.* (PDT)

Although$_{\textbf{concession}}$ *the football season is in full swing, Sparta players have not yet received a premium for winning the championship last year.*

If the results presented in Table 9.7 are taken into account, we can see that *concession* occupies a position which is closer to *reason–result* than to *opposition*. We can thus claim that, from a formal point of view, *concession* is related rather to CONTINGENCY relations than to CONTRAST class.

Explication has a mixed nature of a very similar kind – on the one hand, it is related to relations as *reason–result* because some fact present in the first argument is typically justified in the second argument (the relation thus belongs to the CONTINGENCY class). On the other hand, this explanation is given by adding some details about what has already been said, so the relation is related to *specification* (from EXPANSION class), too. In Example 149, the clause after the dash explains not only why the atmosphere was recalled but also that it was recalled and how it was done (there is a typical connective for *explication* in Czech – connective *totiž* – which has no exact translation into English, it could be roughly approximated by the expressions *in fact, actually*, we thus leave it without translation in the presented context).

(149) *Atmosféru někdejších zápasů federální reprezentace paradoxně připomněly úvodní hymny – nejprve* ***totiž***$_{\textbf{explication}}$ *na počest hostujícího týmu zazněla česká a po kratičké pauze slovenská.* (PDT)

The atmosphere of former Czechoslovak matches was paradoxically evoked by the introductory anthems –$_{\textbf{explication}}$ *in honor of the visiting team* [*totiž*]*, first the Czech anthem and after a short pause the Slovak one were played.*

In this case, the characteristics presented in Table 9.7 seems to confirm the mixed nature of the relation – *explication* stands between *reason–result* and *specification* resembling the latter more than the former.

9.3 Summary

In this case study, we focused on the relation between discourse and syntactic analysis of the text following our experience with discourse annotation of Czech data in the Prague Dependency Treebank. This relation was discussed from two different perspectives – in the first part, the study showed which syntactic characteristics were exploited during the discourse analysis of the text, the second part, on the contrary, took discourse relations as the starting point and commented their characteristics from the syntactic perspective.

In the first part of this study, we found that for all aspects of discourse analysis there were some corresponding syntactic characteristics marked at the tectogrammatical layer of the PDT. To some extent, discourse semantic types of relations can be automatically detected using syntactico-semantic labels from the tectogrammatical layer, the scope of arguments is in most cases indicated by the tree structure and connectives can be found at pre-specified places in trees or are signalled by specific labels. The most distinctive evidence of correspondence between syntactic and discourse-level analyses is the fact that almost 80% of discourse relations realized within a single sentence were extracted automatically using syntactic features and that syntactic annotation defines the scope of discourse argument properly in 99% of cases. Furthermore, the coordination resolution makes discourse annotation more comprehensible and ellipsis resolution enables the annotation of discourse relations even in cases with elided verbs in the surface form of sentence.

On the other hand, discourse analysis differs significantly from the syntactic analysis – first of all, unlike syntactic analysis, it goes systematically beyond the sentence boundary (and thus syntactic features do not offer any clues for it) and moreover, it puts greater emphasis on meaning rather than on the formal aspect of expressions. Therefore, some mismatches between syntactic and discourse analyses are inevitable. We mentioned cases where the syntactic analysis approaches some structures either according to their form (structures with apposition, some relative clauses, clauses with connective *along with* and so on) or only from a very general semantic perspective (*Adversative* structures). We also showed cases where the syntactico-semantic analysis differs significantly from the discourse point of view. Further, it should be noted that automatic extraction of discourse relations using syntactic features could take place only after manual analysis – in other words, annotators indicated places where syntactic analysis corresponds to discourse-level analysis by marking places where it is not the case.

The second part of the study provided some observations on the syntactic characteristics of discourse relations. For all findings, it is necessary to bear in mind that the concerned analysis only deals with relations signalled by overtly present connectives. We found that intra-sentential relations are more numerous than relations between separate sentences (approximately 12,600 intra-sentential versus 5,500 inter-sentential). Among relations realized within a single sentence, coordinate structures are predominant (70% of all intra-sentential realization), while subordinate constructions represent discourse relations rather marginally.

Concerning the syntactic realizations of each semantic discourse type separately, there are certain scales for both the intra- versus inter-sentential forms of expression and for subordinate versus coordinate structures. While semantic types of *condition*, *purpose*, *disjunctive alternative* and *pragmatic condition* occur in our data in more than 90% of cases within a single sentence, only a single semantic type, *generalization*, has such a high percentage of inter-sentential realizations. Other semantic types are placed between these two poles, the majority of them having the predominant realization within a single sentence. As for the proportion of subordinate and coordinate structure, we saw that while the majority of semantic types is realized preferably as coordinate structures, only three of them (*pragmatic condition*, *condition*, *purpose*) are realized in more than 90% as subordinate structures.

To our best knowledge, this case study presents the first complex corpus-based attempt to compare syntactic and discourse analysis for the same data. Although it is for now limited to Czech and a single syntactic theory, it offers several areas for possible comparison with other languages, especially in terms of general syntactic concepts such as relations within a single sentence versus between more sentences, relations in coordinate versus subordinate structures or discourse relations in structures with relative clauses.

10

Morphosyntactic Characteristics of Czech Connectives

As can be seen in all theoretical chapters of this monograph, many different means are engaged in ensuring discourse coherence. In this chapter, we take a closer look at one group of these means – at discourse connectives.[104] Their basic role in ensuring discourse coherence comprises of signaling the relation between individual text spans and expressing the semantics of such a relation. To describe connective properties, many different perspectives can be adopted. This chapter focuses on syntactic and morphological characteristics of Czech connectives applying two different perspectives. First, basic information about connectives from relevant literature are summarized and problematic areas of this traditional description are mentioned. At the same time, we comment these characteristics from the point of view of discourse level analysis in the PDT, where necessary. In the second part, most frequent connectives from the PDT 3.0 are characterized in detail. A quantitative analysis of all the annotated data (49,431 sentences) enables us to accompany the properties described in general in the first part of this chapter with authentic examples.

10.1 General Characteristics

Having observed all types of devices with a discourse-structuring function and having searched for a delimitation of the category of discourse connectives, we came up with three general groups of these language means. The most common group is represented by word units, then there are numbers and letters that can function as discourse connectives and, as a final group, some punctuation marks seem to have the text-structuring function, as well. In this study, we leave the numbers and letters aside, since they function as connectives mainly in list structures and their other usage is restricted only to conjunction. As for punctuation marks, comma and full stop were found to be too indistinct in their meaning for discourse level analysis and are therefore not considered to be discourse connectives. On the other hand, colon, semicolon and dash can very often signal some discourse-relevant meaning, mostly some type of adordination (apart from other functions, e.g. for colon, signaling direct speech). Both colon and dash belong to the most frequent connectives in the PDT,

[104] This chapter targets only the *primary connectives*. The less frequent and morphosyntactically different group of *secondary connectives*, or *multiword discourse phrases*, is the topic of Chapter 11.

their function is therefore discussed in the second part of this chapter. The rest of the chapter is devoted exclusively to word connectives.

10.1.1 Part-of-speech classification

A general, functional definition of discourse connectives was given earlier in Chapter 2. Discourse connectives were characterized as language expressions whose function is to join or connect pieces of discourse to a meaningful whole while at the same time signaling a semantic relation between them.

A closer look at these expressions reveals that they belong to many part-of-speech (PoS) categories. If all such lexical units are taken into account (including secondary connectives, cf. Chapter 11), we find that the only PoS category that is not involved in the construction of expressions with the discourse-connecting function are interjections.

In this chapter, a narrower view of connectives is applied – only morphologically stable, mostly non-declinable expressions are further discussed. In Czech linguistics, the most common connectives are classified as conjunctions, adverbs and particles. Whereas the connective function is primary to conjunctions, other connectives have also other functions in texts.[105] In some approaches, PoS classification depends on the function of the lexical unit (e.g. the lexical unit *však* [*however*] is considered to be a conjunction or a particle according to its function in a text), sometimes only prevalent function is taken into account for classification. There is also an approach which divides connectives according to their intra-sentential versus inter-sentential usage (Hrbáček, 1994), but this study does not work with this feature as a primary criterion.

Concerning criteria for PoS classification, so far one of the most respected Czech grammar book (Komárek et al., 1986) comments on the PoS properties of Czech connectives rather marginally. It claims that conjunctions are expressions which signal relations of subordination and coordination. Expressions which convey the meaning of adordination are considered to be particles. However, this semantic characteristics remains without any further details.

More elaborated attempts to characterize the PoS properties of Czech connectives can be found in some studies only on particular connectives. The following paragraphs give a summary of the criteria used for deciding on the PoS appurtenance of the examined connectives.

In a simplified view, the PoS appurtenance of an expression depends on its function in the text: Conjunctions are expressions with connective functions whereas non-connective function is treated as a basic feature of particles.

In more elaborated studies, conjunctions are viewed as expressions with obligatory initial position in a sentence (Pešek, 2011) which do not influence the position of clitics

[105] Similarly, Stede and Neumann (2014) note that 40% of German connectives also have a non-connective reading.

in the sentence (ibid.), they cannot be rhematized (ibid.) and do not combine with the conjunction *a* [*and*] (based on the assumption that a conjunction cannot be combined with another conjunction, cf. Bauer, 1972).

Adverbs are discourse-structuring devices which can be combined with the typical conjunction *a* [*and*]; they can be rhematized, i.e. they can represent the focus proper (Pešek, 2011) and often have a certain coreference relation to the surrounding context (Bauer, 1972).

Particles are treated as discourse-structuring devices only marginally. The only feature we came across in literature is the presence of presupposition at some of them (Bedřichová, 2008).

Thus, we can conclude that a detailed part-of-speech characteristic of connectives does not exist so far in Czech. There are rather comments on properties of connectives than a complete study. This case study cannot, for obvious reasons, offer any kind of integrated view of all connectives either. Rather, it aims to point out the possibilities of connective characterization which arose during the annotation of discourse relations in the PDT. This starting point is a great advantage, since it represents the most complete attempt to capture and observe connectives in Czech so far. On the other hand there is also a slight limitation – non-connective uses of the expressions in question cannot be studied easily using this annotation. Being aware of this, we concentrate on the issues of such characteristics of discourse connectives in the second part of this chapter and forego an integrating PoS approach towards this group.[106]

10.1.2 Form and inflection

One of the most common properties of Czech connectives is their indeclinability. However, there are also three connectives which differ in this respect. First, the expressions *aby* (roughly [*in order to*] in purpose constructions in English) and *kdyby* [*if*] function as connectives and as parts of a verbal form at the same time: A conditional verb form in Czech is created by an auxiliary form of the verb *být* [*to be*] (i.e. one of forms *bych*, *bys*, *bychom*, *byste*, *by*) and an active participle of main verb (e.g. *řekl* is an active participle of verb *říci* [*to say*]). The whole form is thus, for example, *by řekl* [(*he*) *would say*]. The connectives *aby* [*in order to*] and *kdyby* [*if*] arose as a combination of this special conditional form of the verb *to be* and connectives *a* [*and*] (*a+by > aby*) or *kdy* [*when*] (*kdy+by > kdyby*). The morphological characteristics of the verbal parts are preserved, so the connective is conjugated according to person and number as a regular verb, cf. *aby řekl* [*in_order_that_he would_say*], *abychom řekli* [*in_order_that_we would_say*], *kdyby řekli* [*if_they would_say*], *kdybyste řekli* [*if_you_pl would_say*] etc. Example 150 illustrates the connective *kdyby* in first plural form *kdybychom*, Example 151 shows the connective *aby* in second plural form *abyste*.

[106] Another solid reason for this resignation is the traditional lack of clear and testable criteria for distinguishing among functional (or synsemantic) PoS categories.

(150) ***Kdybychom** se nechali porazit jako ostatní, nemuseli jsme potom dodatečně před světem dokazovat rezistenci kupříkladu atentátem na Heydricha.* (PDT)

__If__ we had let them beat us like the others, then we wouldn't have to prove subsequently our resistance to the world for example by the assassination of Heydrich.

(151) ***Abyste** to překonal, chce to vůli a trpělivost, prostě to nevzdat.* (PDT)

***To** overcome it* (lit. *in_order_that_you it would_overcome*), *you need will and patience, simply do not give up.*

Second, there is the connective *což* [*which*], which arose from a combination of a relative pronoun *co* [*what, which, that*] and a bound particle *-ž*. If this pronoun refers to a whole clause, the connective can be roughly paraphrased by *and this* (for details about coreference and the connective function see Chapter 9). The connective *což* is declined as a regular pronoun and as such it is also a participant in the clause structure. Consider, for example, sentences 152–154: In the first of them the connective *což* is in the nominative case (*což*) and has the role of subject. In the second example, it is in the instrumental (*čímž*) and stands at the position of object. In the third sentence, the pronoun is in the genitive (*čehož*) and functions as an adverbial complement.

(152) *V letech 1994–1995 musí podnik zřídit kontinuální měření emisí, **což** přijde na 50 milionů Kč.* (PDT)

*In 1994–1995, the company must establish a continuous measurement of emissions, **which** comes to 50 million CZK.*

(153) *Například naše zubní pasty obsadily dominantní podíl 55 procent, **čímž** se nemůže pochlubit ani žádná světová firma.* (PDT)

*For example, our toothpaste occupied a dominant share of 55 percent, **which** is not the case even for any global company* (lit. ***by which** no global company can boast*).

(154) *Ve škole se nudí, zlobí, na základě **čehož** od některých učitelů dostávají špatné známky.* (PDT)

*They are bored in school, they are disruptive, **which** leads some teachers to give them poor grades* (lit. *on the basis of **which** they receive poor grades from some teachers*).

The connective *což* [*what, which, that*] also gave rise to other Czech connectives when combined with prepositions. For example, the temporal connective *přičemž* [*whereas*, lit. *by_which*] is composed of the preposition *při* [*by*] and the locative case of the pronoun *co* [*what*], i.e. the form *čem*, and the bound particle *-ž*. Further, the connective *pročež* [*therefore*, lit. *for_which*] is created by the preposition *pro* [*for*] and a shortened version of accusative of the pronoun *co*, i.e. the form *č*, and again the bound particle *-ež*. The forms of these connectives are already completely fixed (they are indeclinable).

However, formal issues related to connectives are not solved by (in)declinability. There is also one more common problem with the formal aspect of Czech connectives, namely what word combinations should be considered a single connective and which represent more individual connectives. Connectives often fortify their meaning mutually, sometimes, especially in case of connectives with a vague meaning like *and* or *but*, other, additional parts modify their meanings substantially. Most probably, we can assume all parts of such a word combination to represent a single connective in all cases where only one semantic type is expressed – even if the connective looks as complicated as *na jedné straně – na druhé straně však* [*on the one hand – on the other hand however*], *nejen – ale především* [*not only – but also especially*] or *i kdyby – přesto* [*even if – still*]. Word combinations signaling more semantic types need to be treated as separate connectives – cf. connectives *ale* [*but*] and *potom* [*then*] in Example 155: The first one signals *opposition*, the second signals *asynchrony* between the same arguments.

(155) *Jemně a s mírou ochutnal nabízené víno,* ***ale*****opposition** ***potom*****asynchrony** *se jako pravý profesionál omluvil, že musí do hotelu zopakovat si úlohu, protože se následující den natáčí poslední scéna.* (PDT)

He tasted the offered wine gently and decently, ***but*****opposition** ***then*****asynchrony** *he apologized like a pro that he had to return to the hotel to repeat his role, for the last scene was to be filmed the next day.*

10.1.3 Origin

The partial transparency of the form of some previously mentioned connectives leads us to the third characteristics of connectives – the internal structure of some of them can be easily analyzed from the perspective of modern language, some of them are now opaque, but their origin can be traced from a historical perspective and, finally, there are connectives whose origin is completely hidden.

Connectives **analyzable from the synchronic perspective** show some regularities which occurred in the course of historical development of the connective (as further illustrated below). Two general groups can be distinguished: The first consists of connectives whose base is created by a conjunction accompanied by some other part – e.g. a particle or another conjunction. For example, the conjunction *neboť* [*because*] arose from the combination of the conjunction *nebo* (nowadays [*or*], historically [*because*] as well) and the particle *-ť*. The conjunction *avšak* [*however*], orig. lit. [*and_but*], is evidently traceable to the conjunctions *a* [*and*] and *však* [*but*]. A similar case represents the conjunction *anebo* [*or*], orig. lit. [*and_or*], consisting of the conjunctions *a* [*and*] and *nebo* [*or*]. Further, the connective *ačkoli* [*though*], orig. lit. [*though_ever*], can be analyzed as the conjunction *ač* [*though*][107] and the particle *-koli* (roughly [*ever*]) etc.

[107] Whereas from the synchronic perspective, the conjunction *ač* is opaque, the diachronic view reveals that it arose from a combination of the conjunction *a* [*and*] and the extinct particle *če* signaling surprise (Bauer, 1960).

A second group comprises connectives with the structure preposition + demonstrative pronoun (+ conjunction). Consider for example the connectives *zatímco* [*while*], *protože* [*because*] or *přestože* [*though*]. The first connective, *zatímco* [*while*], orig. lit. [*as_this_that*], is put together from preposition *za* [*during*], as a second component there is a form of the pronoun *to* [*this*] (in the instrumental form *tím*) and the last part is represented by the conjunction *co* (originally [*what*]).[108] The connective *protože* [*because*], orig. lit. [*for_this_that*], consists of the preposition *pro* [*for*], the accusative form of the pronoun *to* [*this*] and the conjunction *že* [*that*]. Similarly, the connective *přestože* [*though*], orig. lit. [*over_this_that*], is composed of the preposition *přes* [*over*] and the same forms of the pronoun *to* [*this*] as in the previous example and the conjunction *že* [*that*].

There are also connectives compounded from a preposition and a pronoun only, like *proto* [*therefore*], *přesto* [*though*] or *přitom* [*while*, *yet*]. The connective *proto* [*therefore*], orig. lit. [*for_this*], is composed of the preposition *pro* [*for*] and the pronoun *to* [*this*] in the accusative. The connective *přesto* [*though*], orig. lit. [*over_this*], is created by the preposition *přes* [*over*] and the pronoun *to* [*this*] in accusative. The connective *přitom* [*while*, *yet*], orig. lit. [*during_this*], by the preposition *při* [*by*] and the pronoun *to* [*this*] in locative (the form *tom*). Connectives like *kvůli tomu* [*because of that*], orig. lit. [*because_of that*], or *mimo to* [*apart from that*], orig. lit. [*apart_from that*], clearly indicate that this process is still alive in modern Czech.[109]

Apart from these two general groups, there are also cases with unique origin which are analyzable from the perspective of modern language: e.g. the connective *třebaže* [*although*], orig. lit. [*maybe_that*], is from the synchronic perspective analyzable as the adverb *třeba* (roughly [*maybe*, *even*]) and the conjunction *že* [*that*]. The connective *když* [*when*] is traceable to the adverb *kdy* [*when*] and the bound particle *-ž*. Similarly, the connective *naopak* [*on the contrary*], orig. lit. [*on_opposite*], would be analyzed as the preposition *na* [*on*] and the noun *opak* [*opposite*].

Then, there are connectives **opaque from the synchronic perspective but their origin is traceable from a historic point of view** – they are very diverse.[110] Some of them evolved from autosemantic adverbs – e.g. the connective *tedy* [*thus*] originally had the meaning *in that time*, the connective *však* [*however*] was used with the meaning *in every way*. Other connectives arose from the combination of simpler conjunctions or particles, e.g. the connective *nebo* [*or*] can be traced to the negative particle or adverb *ne* and particle *bo*, one of the most frequent connectives *ale* [*but*] developed from the combination of the conjunctions *a* [*and*] and *le* [*and*, *but*] etc. The most fascinating deve-

[108] The expression *co* [*what*] is originally a pronoun but it gave rise also to the conjunction *co* in modern Czech.

[109] The structure of these connectives allowed us to use textual coreference for their identification in texts – the pronominal part relates to the preceding context with a coreference link (more details can be found in Rysová and Mírovský, 2014b and Poláková, Jínová and Mírovský, 2012).

[110] For information on these connectives, we use Dictionary of Old Czech (Gebauer, 1970) and Czech Etymological Dictionary (Rejzek, 2004).

lopment is in our opinion represented by the connectives *totiž* (without equivalent in English, roughly translatable as *actually*, *in fact* or *that´s to say*) and by the first part of the connective *buď – anebo* [*either – or*]. The connective *totiž* developed as a fusion of the second person singular present form of the verb *to perceive* (in Czech *čuješ* and afterwards *čúš*) and the object of this verb *it* (in Czech *to*), so it is originally lit. *to-čúš* [*it-you_perceive*]. The resulting form probably gets fused with the older particle *totiť* with a similar meaning. Its meaning in modern Czech evokes this origin – it serves mostly for explanations in argumentation. In the second case, the first part of the connective *either – or buď – anebo* evolved through the fossilization of the imperative of the verb *to be*, thus the original literary meaning of the connective *buď – anebo* is *be_it – and_or*. The imperative form of the verb *to be* has always been *buď* but native speakers of Czech nowadays never connect this verb form with the connective.

Only few connectives have been **opaque** in the whole known history of Czech. We found only two of them – *a* [*and*] and *le* [*and*, *but*], the latter moreover does not function as a separate connective anymore (it is only a part of the connective *ale* [*but*]). This finding indicates that discourse connectives are relatively newly established forms in language which evolved from the need to express more and more complicated contents. They are definitely associated with refined intellectual use of language rather than with everyday spoken communication.

10.1.4 Placement in the sentence and in the argument

The fourth perspective for observing properties of connectives is looking at their placement in a sentence and in a discourse argument. We adopted two different starting points for observing placement of connectives in a sentence. First, applying the word order perspective, we can distinguish connectives with placement in the first, the second or in either of these positions and connectives whose placement depends on the topic–focus articulation of a sentence. For example, all subordinate connectives (like *protože* [*because*], *když* [*when*], *pokud* [*if*]) take, in prototypical cases, the first position in the sentence. Also, some coordinate connectives are placed in the first position in the second clause in a compound sentence (this is the case for connectives like *nebo* [*or*] or *takže* [*so*]).

On the other hand, some connectives almost obligatorily take the second position in a sentence if they express a certain meaning, and the first position if they express another meaning – e.g. the connectives *totiž* (roughly [*actually*, *in fact*]) and *tedy* [*thus*] are almost always placed in the second position in a sentence if they express *reason–result* or *explication*, but they can stand in the first position if they are used for other meanings. Further, the connective *však* [*however*] is obligatorily placed in the second (clitic) position if it signals a contrastive meaning, but, if it is used for other meanings or in a non-connective function, it is placed at the beginning of the clause.

There are also connectives which can be placed in the first or second positions independent of their meaning – e.g. *navíc* [*moreover*], *proto* [*therefore*], *dále* [*further*].

This characteristics is sometimes connected with the origin of the connectives – we can observe that connectives with a deictic component as their ancestor often take the second position in a sentence, i.e. the clitic position which is connected with short forms of pronouns in Czech.[111] On the other hand, the prototypical placement of a connective is at the beginning of the sentence, so these two tendencies (the origin and prototypical placement of a connective) are contradictory and lead to variation.

The only group of connectives whose placement depends on topic–focus articulation are rhematizing particles like *také* [*also*] or *jen* [*only*]. As was already mentioned in Chapter 9, rhematizing particles were considered to function as discourse connectives only if they were connecting two different verbal groups (not mere nominal groups).

Another perspective in the discussion on connective placement takes into account the order of the discourse arguments and their participation in the expressed meaning. There are connectives which occur only in one argument (e.g. the connective *však* [*however*] is always placed in the second argument), then connectives which are placed (or can be placed) in both arguments (e.g. connectives with more parts like *buď – anebo* [*either – or*], *sice – ale* [*albeit – however*], *jestliže – pak* [*if – then*]) and also connectives whose placement depends on the meaning expressed in the given argument. The last characteristics is typical for connectives of discourse relation *reason–result* – there is a group of connectives signaling reason (e.g. *protože* [*because*], *neboť* [*because*]) and on the contrary a group of connectives expressing result (*proto* [*therefore*], *tedy* [*thus*], *takže* [*so*]).

10.1.5 Subordinate, coordinate and inter-sentential connectives

The last comment on general properties of discourse connectives concerns their syntactic characteristics. In Czech grammar, there is a long tradition of distinguishing between subordinate (e.g. *když* [*when*], *pokud* [*if*]) and coordinate conjunctions (e.g. *ale* [*but*], *takže* [*so*]). The main difference lies in their placement – whereas subordinate connectives are a part of the dependent clause (which can take the first position in the compound sentence), coordinate connectives are always placed between clauses. The second difference between coordinate and subordinate structure is the fact that a compound sentence with a subordinate conjunction cannot be reformulated into two separate sentences preserving the given subordinator, while there is almost always this possibility in case of compound sentences with a coordinate conjunction. So far, the Czech grammar disregards inter-sentential connectives or it states that these connectives act similarly as coordinate conjunctions. From our perspective, we can add that (i) coordinate conjunctions function often as inter-sentential connectives but we found one coordinate connective for which it is not the case – the connective *nýbrž* [*but*] is used intra-sententially only – and (ii) all connectives other than conjunctions can participate in the construction of coordinate compound sentences as well.

[111] For the distinction between short and long forms of pronouns see Chapter 5, Section 5.4.1.

10.2 Characteristics of Most Frequent Connectives

The current section gives an overview of twenty most frequent primary connectives as they were annotated in the texts of the PDT 3.0 and it describes their morphosyntactic properties. For the first time in the Czech corpus linguistics, we are able to document various properties of these expressions empirically on a large amount of authentic texts.[112]

In this way, we can verify basic assumptions about the functionally delimited category of discourse connectives and also find evidence for their new, unexpected properties.

From the many morphosyntactic aspects, we focus here on the PoS distribution in the core of the category (10.2.2), on the intra- and inter-sentential use of these connectives (10.2.3) and on the distribution of the core of the group as discourse connectives and in other, non-connective uses, learning in this way about the prevalence of the discourse-connecting function for the given expressions (10.2.4).

10.2.1 Frequency

First, we introduce a list of twenty most frequent Czech connectives in the PDT 3.0 data. The list was obtained by a simple search for the most frequent forms with the function of a discourse connective. The frequency figures refer to frequencies of forms of connectives (e.g. *Ale* [*But*] is included in *ale* [*but*]). Connectives occurring in list structures are disregarded (they are non-typical – numbers, letters, stars etc.). The findings relate to the given forms alone (e.g. *aby* [*in order to*] does not include the occurrences of *abyste* [*in_order_that_you*] or *abychom* [*in_order_that_we*], *nebo* [*or*] does not include the two-part connective *buď – nebo* [*either – or*]).

The list of expressions representing the top-twenty discourse connectives in our data is not surprising per se, cf. Table 10.1. It covers mainly regular and frequent Czech expressions used across domains with a connecting function within a sentence or between sentences. The conjunctions *a* [*and*] and *však, ale* [*but*] lead the list. Not expected, however, are some minor observations: If we allow some punctuation marks to be treated as connective devices, they become one of the most frequent ones (colon and dash, cf. Table 10.1). This is probably because, while occurring very frequently, they also have the potential to express many meanings. For instance, disambiguating appositions (of two verb containing structures, cf. Example 156) was one of the tasks where lots of colons and dashes revealed to have different discourse semantic functions. In Example 156, the colon was annotated as a connective with the discourse meaning of *specification*.

[112] We are aware of the obvious limitation due to the nature of our data: we primarily document the use of connecting devices in written contemporary Czech within the journalistic domain.

Connective	Number of occurrences	PoS	Intra-sentential	(%)
a [*and*]	5,820	JCon	5,477	94
však [*but, however*]	1,527	JCon	266	17
ale [*but*]	1,275	JCon	850	67
když [*when*]	575	JSub	575	100
protože [*because*]	525	JSub	518	99
totiž [*actually, in fact*]	461	Db	24	5
pokud [*if*]	404	JSub	404	100
:	396	Z	349	88
proto [*therefore*]	380	JCon, Db	41	11
tedy [*thus, so*]	308	Db	33	11
aby [*in order to*]	306	JSub	305	99
pak [*then*]	296	Db	78	26
ovšem [*yet, though*]	293	JCon, TT	66	23
-li [*if*]	249	TT	249	100
také [*also*]	234	Db	9	4
neboť [*because*]	221	JCon	220	99
–	218	Z	216	99
zatímco [*while*]	204	JSub	203	99
nebo [*or*]	191	JCon	167	87
což [*which*]	189	PE, TT	184	97

Table 10.1: Twenty most frequent connectives in the PDT 3.0, their PoS and intra-/inter-sentential use. The PoS tags mean: JCon – Coordinate conjunction (connecting main clauses, not subordinate), JSub – Subordinate conjunction (including *aby* [*in order to*], *kdyby* [*if, in case*] in all forms), Db – Adverb (without a possibility to form negation and degrees of comparison, e.g. *pozadu* [*backwards, behind*], *naplocho* [*flatly*]), Z – Punctuation, PE – Relative Pronoun, relative pronoun *což* (corresponding to English *which* in subordinate clauses referring to a part of the preceding text), TT – Particle.

(156) *Spisovatelovo umění se nezapře:*$_{\textbf{specification}}$ *málokomu se podaří vtěsnat tolik nenávisti a lži do jedné věty.* (PDT)

The writer's art cannot be denied:$_{\textbf{specification}}$ *Very few manage to squeeze such an amount of hatred and lies into one sentence.*

Also surprisingly, the relative pronoun *což* [*which*], once decided to be treated as a connective, got to the twentieth position in the frequency chart, even though its other inflected forms were not included in the measurement.

10.2.2 Part-of-speech characteristics

The third column of Table 10.1 presents the part-of-speech appurtenance for each of the frequent connectives. More precisely, it gives the first two positions of the morphological tag used in the analysis on the morphological level[113] of the PDT (Hajič, 2004; Hana et al., 2005). The legend for the tag abbreviations (slightly modified for easy understanding) follows the table.

As Table 10.1 indicates, there are twelve conjunctions (seven coordinate and five subordinate) among the twenty most frequent connective types. This proportion goes hand in hand with the previous claim that the connecting function primary for conjunctions (in sentential analysis) makes this PoS category also "primary connectives," in other words, it makes them expressions with the highest potential to also interconnect discourse segments. Four connective types belong to adverbs, two are punctuation marks. One of them is a conditional particle *-li*, which cannot stand on its own – it only occurs as a clitic connected typically to the first word in an utterance. Only three expressions in Table 10.1 can have more possible PoS categories in function of discourse connectives: *proto* (conjunction and adverb), *ovšem* (conjunction and particle) and *což* (relative pronoun and particle).[114] If we then look at the PoS values for the non-connective uses of these twenty expressions, they are, quite unexpectedly, not richer, or more ambiguous: With the exception of the connective *což*, which can also marginally function as an interjection *nu což*, roughly [*oh well*], it is exactly the same PoS categories within and outside the connective group.

A closer look at the PoS tags in Table 10.1 reveals, however, that other categorization of the listed expressions would be possible in Czech. Although the accuracy of the morphological tagging in the PDT reached approx. 95% (see Chapter 7), discourse connectives represent exactly those cases, where a clear PoS characteristic is a difficult task, as was already indicated earlier in Section 10.1.1. Formal criteria applied to distinctions among other PoS categories cannot be used here (indeclinability, no participation in the sentence structure etc.). According to MorfFlex, a recently released morphological dictionary for Czech (Hajič and Hlaváčová, 2013),[115] seven of the expressions in Table 10.1 have more than one PoS characteristics: *však*, *totiž*, *proto*, *tedy*, *aby*, *ovšem*, *což*. (Again, there are only three such cases in the PDT tagging.) A comparison of the PoS assignment in the PDT and the possible PoS values for the expressions in question in MorfFlex manifests two tendencies: (i) the PoS categories for Czech expressions constituting frequent discourse connectives undergo a finer disambiguation, and (ii) this disambiguation is obviously function-based. For instance, the expression *tedy* [*thus*, *so*] has a single PoS category throughout the PDT (adverb)

[113] manual analysis with semi-automatic checks

[114] However, there is only a single occurrence of the particle *což* – in the meaning of *Což jsem to neříkal?*, roughly [*Did I not say it?*].

[115] MorfFlex is primarily based on the Dictionary of Standard Czech (Havránek et al., 1989), further enriched by other literature and continuously maintained.

whereas it has two possible PoS categories in MorfFlex which are conjunction and particle (no adverb).

10.2.3 Intra- and inter-sentential use of connectives

The intra- and inter-sentential distribution of discourse relations and their individual semantic types in the PDT 3.0 was described in Section 9.2. In this section, the intra- and inter-sentential uses for the most frequent connectives are presented. The fourth column of Table 10.1 gives the absolute number of intra-sentential uses of the connectives and the percentage. These figures are particularly interesting in light of the considerations about PoS appurtenance above. The percentage of intra-sentential use approaches 100% in case of subordinate conjunctions – as expected, individual exceptions are dependent clauses with no governing clauses, it is the cases of segmentation of a dependent clause,[116] cf. Example 157:

(157) *Kdo půdu udržel, stal se praotcem šlechtického rodu.* ***Protože*** *kdo držel půdu, byl svobodný.* (PDT)

Who retained the land became the forefather of an aristocratic family. ***Because*** *who held the land was free.*

Adverbs are, on the contrary, used with the same consistency predominantly inter-sententially. The proportions are very dispersed for coordinate conjunctions: They range from 10.8% for *proto* [*therefore*] – an expression that sometimes borders conjunction and adverb readings, across 66.7% for *ale* [*but*], which is one of the most typical Czech coordinate conjunctions, to 99.6% for *neboť* [*because*]. Notable is also the fact that *a* [*and*] occurs almost exclusively intra-sententially (94.1%) whereas *však* [*but, however*] has a strong inter-sentential tendency (only 17.4% of intra-sentential uses), and so it behaves more like an adverb. This varying degree of ability to connect separate sentences and larger discourse units (as far as we could document it in our data) is very likely connected with two factors: First, it is the rules of placement of a connective (conjunction) in a sentence explained above in Section 10.1.4. The expression less fixed to a certain position within a sentence can "move more freely" and so is better capable of creating long-distance connections, in this case, across the sentence boundary. The second factor concerns connectives containing a referential component (even if it is not that apparent from their historical development, cf. Section 10.1.3 above): Thanks to the referential ability of this morpheme, these connectives can relate, similarly as demonstrative pronouns, to distant segments of texts (cf. the analysis of connectives with a referential component in Poláková, Jínová and Mírovský, 2012).

[116] in the Czech linguistic tradition known also as *parceling*

10.2.4 Degree of connectivity

Apart from the PoS categorization of frequent discourse connectives and their intra- and inter-sentential use, another property is worth addressing in this section: their degree of connectivity. It is the proportion of connective and "non-connective" occurrences of the given forms in our data. "Non-connective occurrences" refers to the difference of the total number of occurrences of the given expression in the PDT (or more precisely, in the part annotated for tectogrammatics and discourse) and such uses, where the given expression is annotated either as a separate connective (e.g. *však*) or as a part of a connective (*přesto však*). The results are presented in Table 10.2. The degree of connectivity, in Table 10.2 represented as the percentage of discourse connective (DC) use in the PDT data, clearly correlates with other functions (and possible other PoS characteristics) of the expressions in question. This is expectable and natural; what is interesting are the proportions for the individual expressions. For example, *však* [*but, however*], *tedy* [*thus, so*] and *ovšem* [*yet, though*] in Czech are known to also function as expressive particles. This function of the expression *však* is demonstrated in Example 158. For comparison, in Example 159, the same expression has the function of a discourse connective and signals the discourse meaning of *opposition*.

(158) ***Však*** *ony se ještě budou hodit, až se něco zadrhne.* (PDT)
They can still be useful ***after all****, when something goes wrong.*

(159) *Doklady k odpočtu se k přiznání nepřikládají. Musíte je* ***však*****opposition** *uchovávat pro případ kontroly po dobu 10 let.* (PDT)
Documents for the tax deduction are not to be attached to the tax return form. ***However*****opposition***, you must keep them available for inspection for the next 10 years.*

However, the degree of connectivity for these three expressions varies a lot: Whereas *však* [*but, however*] functions as a connective in almost 94% and *ovšem* [*yet, though*] in 86% of their occurrences, it is only about 58% for *tedy* [*thus, so*]. Other factors may influence these figures, like the tendency of these expressions to relate to an unspecific portion of the preceding text (appearing in questions of dialogs) or to non-expressed contents. Also, the nature of the conjoined parts (abstract objects or entities) is in play in degree of connectivity. This is clearly visible for the most frequent *a* [*and*] and for *nebo* [*or*]: Only a small fraction of them (37.5% and 22.4%, respectively) function as discourse connectives, most likely because they very often relate entities irrelevant for discourse analysis, or nominalizations of abstract objects that have not been annotated as discourse arguments so far.

Table 10.2 further demonstrates that for subordinate conjunctions the degree of connectivity is uniform and high; it ranges from 78% to 88%, with the exception of *aby* (31.3%). This expression, as the only one on the list, has the ability to also introduce content clauses (mostly object clauses after verbs of saying) – these connections are

Expression	Number of occurrences	DC occurrences	Non-DC occurrences	DC use (%)
a [*and*]	17,756	5,820 (+839)	11,097	38
však [*but, however*]	1,774	1,527 (+139)	108	94
ale [*but*]	2,250	1,275 (+466)	509	77
když [*when*]	989	575 (+205)	209	79
protože [*because*]	625	525 (+3)	97	85
totiž [*actually, in fact*]	514	461 (+24)	29	95
pokud [*if*]	571	404 (+69)	98	83
:	2,297	396 (+21)	1,880	18
proto [*therefore*]	654	380 (+232)	42	94
tedy [*thus, so*]	576	308 (+28)	240	58
aby [*in order to*]	1,298	306 (+100)	892	31
pak [*then*]	546	296 (+140)	110	80
ovšem [*yet, though*]	373	293 (+28)	52	86
-li [*if*]	371	249 (+47)	75	80
také [*also*]	1,028	234 (+92)	702	32
neboť [*because*]	225	221 (+0)	4	98
–	2,300	218 (+42)	2,040	11
zatímco [*while*]	233	204 (+2)	27	88
nebo [*or*]	1,028	191 (+39)	798	22
což [*which*]	350	189 (+8)	153	56

Table 10.2: Connective (DC) and non-connective (Non-DC) uses of polysemous expressions in the PDT 3.0. The first number in the third column is frequency of the given form occurring alone as a connective, the second number (in parentheses) is its occurrence as part of a connective.

also outside the scope of our discourse analysis. Once the annotation of discourse arguments covers also nominalizations of abstract objects, the proportion of connective use of conjunctions is expected to rise significantly for the coordinate ones, but not for subordinators: Syntactically they relate mostly to verbal structures.[117]

The highest degree of connectivity show (given the existing annotation so far) the expressions *neboť* [*because*] (98.2%) and *totiž*, roughly [*actually, in fact*] (94.4%), which makes them the prototypical connectives. Looking at the other end of the scale, the punctuation marks colon (18.2%) and dash (11.3%) are not only frequent and polyfunctional as connectives, they are also very frequent (and most likely polyfunctional) outside the discourse connective category.

[117] Structures with subordinators and no verbs like *the emotions are strong because contemporary* or *pokud možno* [*if possible*] are fixed phrases or rather rare.

10.3 Summary

In this chapter, we have outlined the basic morphosyntactic properties of Czech connectives, both from a theoretical perspective and from the perspective of manually annotated data. We have pointed out the issues of the delimitation of the group in an inflective language that Czech represents. We have shown how the richness of forms, their historical development, their morphosyntactic functions (part-of-speech class, degree of connectivity etc.) relate to the shape of the category.

In Czech, the vast majority of primary connectives belongs to non-declinable PoS categories (conjunctions, adverbs, particles) with a few exceptions. Historically, only one Czech connective is opaque in the whole development of Czech (the connective *a* [*and*]); other connectives are traceable all the way to the elementary morphemes, either from the synchronic or the diachronic perspective. Another described property of Czech connectives was their placement in a sentence or in an argument. There are either connectives with a fixed position in the sentence (e.g. subordinators stand obligatorily on the first position in a clause) or with a variable placement (e.g. the conjunction *ale* [*but*] can stand either in the first or second position) and further connectives whose placement depends on the topic–focus articulation of the sentence (i.e. rhematizers). As for the placement in a discourse argument, there are connectives placed only in one argument, connectives present in both arguments (e.g. *either – or*) and also connectives whose placement depends on the meaning expressed in the given argument (this is typical for connectives of the *reason–result* relation).

In the second part of this chapter, we described the morphosyntactic properties for the core of discourse connectives, i.e. for twenty most frequent connectives according to the PDT 3.0 annotation. We have learned that not only declinable discourse connectives, i.e. the relative pronoun *což* [*which*] and the *purpose* marker *aby* [*in order to*], but also the punctuation marks colon and dash made it among the most frequent connectives, and so got way over the frequencies of some basic Czech conjunctions. We have described the proportions of intra- and inter-sentential connective use, finding that coordinate conjunctions behave quite diversely in this respect. We have further verified the claim that the discourse-connecting function is primary for conjunctions by computing the degree of connectivity of individual frequent connective types. With respect to their high frequencies, the prototypical connectives (close to 100% of the occurrences of these expressions in the PDT are indeed annotated as connectives) are *neboť* [*because*] and *totiž*, roughly [*actually*, *in fact*].

The present study of Czech discourse connectives offers first insights into a topic that has not been addressed yet with complexity in Czech. We hope to have shown possible directions for future research both in corpora mining and in the (comparative) research of discourse-structuring devices. The analysis in this chapter concerned the core of discourse connectives, the so-called primary connectives (cf. Rysová and Rysová, 2014). The properties of other connective devices, understood as discourse connectives in a broader concept, is given further in Chapter 11 on multiword discourse phrases.

11

Discourse Relations Expressed by Multiword Discourse Phrases

In Chapter 10, we presented morphosyntactical aspects of Czech primary connectives, i.e. expressions like *a* [*and*], *ale* [*but*], *proto* [*therefore*] etc. that belong to non-declinable parts of speech (mainly to conjunctions, some adverbs or particles). However, during the annotation of authentic Czech texts, we have also encountered many other possibilities of expressing discourse relations, mainly the multiword phrases like *podmínkou bylo* [*the condition was*], *abychom to shrnuli* [*to sum up*], *z tohoto důvodu* [*for this reason*] etc. The aim of this chapter is to introduce and analyze these other possibilities for Czech and to compare them with similar expressions in English.

11.1 A Scale of Explicitness and Implicitness of Discourse Relations

The individual discourse relations may be expressed by specific language means (i.e. explicitly) or they may be implicit (i.e. not signaled by any language expression but only deducible from the meaning of the given discourse units or arguments), see Chapter 2. When dealing with authentic texts, we may observe that the scale between explicitness and implicitness of discourse relations is very rich and extensive, see Example 160:

(160) *Slovenská elita byla zklamána politickou volbou Slovenska.*
Většina kvalitních odborníků zůstala v Praze. (PDT)
Proto *většina kvalitních odborníků zůstala v Praze.*
Z tohoto důvodu *většina kvalitních odborníků zůstala v Praze.*
Kvůli tomu *většina kvalitních odborníků zůstala v Praze.*
Kvůli této skutečnosti *většina kvalitních odborníků zůstala v Praze.*

The Slovak elite was disappointed by the political choice of Slovakia.
Most of the skilled professionals remained in Prague.
Therefore*, most of the skilled professionals remained in Prague.*
For this reason*, most of the skilled professionals remained in Prague.*
Because of this*, most of the skilled professionals remained in Prague.*
Because of this fact*, most of the skilled professionals remained in Prague.*

In this example, there is a discourse relation of *reason–result* between two discourse arguments: *slovenská elita byla zklamána politickou volbou Slovenska* [*the Slovak elite was disappointed by the political choice of Slovakia*] and *většina kvalitních odborníků zůstala* *v Praze* [*most of the skilled professionals remained in Prague*]. We can see that this semantic type of discourse relation may be expressed implicitly or by various language means like *proto* [*therefore*], *z tohoto důvodu* [*for this reason*], *kvůli tomu* [*because of this*] or *kvůli této skutečnosti* [*because of this fact*]. In other words, discourse relations may be signaled not only by the one-word, lexically frozen connectives like *proto* [*therefore*] (i.e. mainly conjunctions, structuring particles etc.), but also by a wide range of other language means, mainly multiword phrases like *z tohoto důvodu* [*for this reason*] etc. These phrases represent an interesting but also difficult class of expressions, as they form a very heterogeneous category in terms of their lexico-syntactic and semantic nature.

On the one hand, these expressions clearly signal discourse relations, on the other hand, they do not belong to the parts of speech generally assumed to be connectives (like conjunctions, some types of particles etc.). The problem with these expressions is that they may be inflected, e.g. *z tohoto důvodu – z těchto důvodů* [*for this reason – for these reasons*] and may occur in many different forms in the text, cf. *kvůli tomu*, *kvůli této skutečnosti*, *kvůli této situaci* [*due to this*, *due to this fact*, *due to this situation*] etc. In this respect, they highly differ from mainly one-word, lexically frozen connectives. They still function, though, as indicators of discourse relations, e.g. the expression *to je důvod, proč* [*that is the reason why*] obviously signals a discourse relation of *reason–result*. Therefore, the discourse analysis or annotation without them would be incomplete. At the same time, as we have already indicated, these expressions are a very heterogeneous class (including prepositional, nominal, verbal phrases etc.). Due to this huge diversity and variability, such expressions are difficult to capture in large corpus data according to some general annotation principles. However, despite these difficulties, the data of the Prague Dependency Treebank (PDT) already contains 1,161 discourse relations signaled by these language means[118] (Rysová and Rysová, 2015) like *výsledkem bylo* [*the result was*], *to kontrastuje s* [*this contrasts with*], *kvůli tomu* [*because of this*] etc.

In the following sections, we describe their lexico-syntactic and semantic nature (and provide a comparison with their English counterparts). The analysis of Czech proceeds from the annotated data of the Prague Dependency Treebank (PDT) and tries to draw a comparison with such expressions in English from the Penn Discourse Treebank (PDTB).

11.2 Terminology: Alternative Lexicalizations of Discourse Connectives vs. Secondary Connectives

In some studies (Prasad, Joshi and Webber, 2010; Prasad, Webber and Joshi, 2014), these expressions or phrases are called *alternative lexicalizations of discourse connectives* (shortly *AltLexes*) or *secondary connectives* (Rysová and Rysová, 2014). AltLexes in

[118] Measured on the whole data of the PDT 3.0 extended with the (yet unpublished) annotation of secondary connectives.

Prasad, Joshi and Webber (2010) and Prasad, Webber and Joshi (2014) are understood as connective expressions that are not defined as explicit connectives in the Penn Discourse Treebank (PDTB). They were discovered during the annotation of implicit relations as places where using a connective from the pre-defined list of English connectives would be redundant. AltLexes are thus a very broad class of language means, containing lexically frozen expressions like *eventually* (that were not included into the pre-defined list of connectives for PDTB annotation), multiword phrases with universal (i.e. context independent) connecting function like *that compares with* as well as multiword phrases that have connecting function only occasionally (i.e. only in certain contexts) like *the increase was due mainly to* (Prasad, Joshi and Webber, 2010).

Secondary connectives (Rysová and Rysová, 2014; Rysová, 2015) is a narrower term – they are expressions with universal[119] (i.e. not context dependent) status of discourse indicators (e.g. *the reason is*, *this contrasts with*, *this was caused by* etc.) that nevertheless differ from the so-called *primary connectives* (e.g. *and*, *but*, *therefore*, *or* etc.) from a lexico-semantic and syntactic perspective, mainly in terms of grammaticalization.

In this case study, we utilize the terminology and definition of discourse connectives given in Rysová and Rysová (2014), which is why we use the term secondary connectives (and when citing previous English studies, we use their terminology, i.e. the term AltLexes).

11.3 Current Annotation of Secondary Connectives in the PDT

As mentioned above, in the current stage, the PDT contains manual annotation of 1,161 secondary connectives, i.e. they form 5.4% of all discourse connectives (both primary and secondary). In this respect, the term secondary seems suitable for these expressions, as their frequency is much lower in the authentic texts than the frequency of primary connectives.

The annotation of secondary connectives was based on the preliminary research in the PDT (Rysová, 2012) carried out on a small sample of data (altogether 261 tokens of secondary connectives). This introductory research was focused on the general characterization of secondary connectives and has opened a unique linguistic topic. We therefore carried out a complete annotation of the whole PDT data, deepening our previous research and providing more detailed conclusions based on extensive linguistic material which we present in this case study.

11.4 Lexico-Syntactic Characteristics of English AltLexes in the PDTB

We begin by looking more closely at the characteristics of English AltLexes in the PDTB which has also inspired our research. Prasad, Joshi and Webber (2010)

[119] "Universal" means that the expression in its connecting meaning may be used in many different contexts to express a given type of discourse relation – e.g. *protože* [*because*] or *důvodem toho je* [*the reason for this is*] are, in this sense, more universal than *důvodem tohoto poklesu je* [*the reason for this decline is*].

evaluate 624 tokens of English AltLexes (manually annotated) in terms of their syntactic and lexical flexibility. First, they examine whether the given expression belongs to one of the syntactic classes admitted for explicit connectives in the PDTB approach (whether it belongs to subordinate conjunctions, coordinate conjunctions, prepositional phrases[120] and adverbs or not). Secondly, they study English AltLexes with respect to their lexical stability, i.e. whether the given expression is lexically frozen or lexically free.

On the basis of these two lexico-syntactic parameters, the authors divide English AltLexes into three categories: (i) syntactically admitted, lexically frozen (*for one thing*, *eventually*); (ii) syntactically free, lexically frozen (*never mind that*, *so what if*); (iii) syntactically and lexically free (*that would follow*, *that is why*). During the PDTB annotation, a list of explicit connectives in English has been introduced; all other expressions with connective discourse function have been annotated as AltLexes. However, the annotation revealed that there are still several expressions that were originally not included in that list of connectives, but that should be there due to their lexico-syntactic nature. In this respect, the authors argue that the annotation of explicit connectives should not be based strictly on a list of expressions, as it will never be fully complete. Thus the expressions of the first group (i.e. syntactically admitted, lexically frozen) should be re-annotated and understood as explicit connectives.

The authors of the study also introduce the basic English patterns for some AltLexes – they argue that AltLexes from the third category are modifiable and therefore may form several different realizations in authentic texts that have the same lexical core plus obligatory and optional elements in the form of noun phrases (NX), prepositional phrases (PPX), verb phrases (VX) or adjectival phrases (JJX).

An example of such AltLex variant and its pattern would be *that may be because* = <*NX*> <*VX*> *because*. We may see that there are some types of AltLexes occurring in the text in different variants or surface realizations (cf. *that may be because*, *this was because*, *this could be because* etc.). A similar situation also occurs in Czech and other languages (like German etc.), cf. several phrases containing the word *důvod* [*reason*]: *to je důvod, proč* [*that is the reason why*], *z tohoto důvodu* [*for this reason*], *důvody jsou různé* [*there are different reasons*] etc. This feature of (syntactically and lexically free) AtlLexes seems to be language general.

11.5 Syntactic Characteristics of Czech Secondary Connectives

In our analysis, we did not focus on whether Czech secondary connectives belong to syntactically admitted classes for connectives or not because there was no predefined list of Czech connectives for practical annotation purposes. The annotators distinguished between primary connectives and secondary connectives themselves during the annotation of authentic texts on the basis of some general instructions

[120] In this chapter, we use the term *phrase* in accordance with the PDTB.

Integrated in the clause structure	Non-integrated in the clause structure (disjunct)
jiný [*different*] *kvůli tomu* [*because of that*] *stejným dechem* [*in the same breath*] *i přes tato fakta* [*despite these facts*] *v důsledku toho* [*as a consequence of this*] *to je důvod proč* [*this is the reason why*]	*jinými slovy* [*in other words*] *krátce řečeno* [*shortly speaking*] *jednoduše řečeno* [*simply speaking*] *přeloženo* [*translated*] *obecně řečeno* [*generally speaking*] *jak je vidět* [*as seen*]
87%	13%

Table 11.1: Czech secondary connectives in terms of their integration into clause structure (examples)

(e.g. by which parts of speech connectives are mostly expressed, how do connectives behave in texts etc.). Moreover, the determination of syntactic classes for explicit connectives is dependent on their general definition that may highly differ according to various individual approaches.

Therefore, in the syntactic characterization of Czech secondary connectives, we concentrate on different issues, particularly on their integration into the clause structure (i.e. whether they function as sentence elements or not) and on their syntactic structure (i.e. whether secondary connectives are mainly verbal phrases, prepositional phrases, whole clauses etc.).

Firstly, we have examined whether Czech secondary connectives are integrated into clause structure as sentence elements or whether they function as clause modifiers (i.e. as the so-called *disjuncts* that are not integrated into the clause structure).

The analysis demonstrated that 87% of Czech secondary connectives (within 1,161 tokens in the PDT) are integrated into clause structure and have a function of sentence elements (e.g. *because of this* is an adverbial of reason) while 13% are clause modifiers commenting on the style or content of the whole clause while remaining unintegrated into its structure as sentence elements (e.g. *simply speaking*),[121] see Table 11.1.

Another examined criterion for Czech secondary connectives is their syntactic structure, i.e. we have analyzed types of syntactic phrases in which secondary connectives appear in authentic texts. The analysis of language material demonstrated that the annotated secondary connectives are realized either by noun phrases, adjectival phrases, numeral phrases, verbal phrases, adverbial phrases, prepositional phrases, particle phrases or by a (semi-)clause (containing either finite or non-finite verbs), see

[121] The analysis does not include secondary connectives in form of the whole, separate units like *důvod je jednoduchý* [*the reason is simple*] that are neither clause elements nor clause modifiers and that stay (syntactically) outside the discourse arguments.

Syntactic phrases	Examples of secondary connectives
noun phrases	*stejným dechem* [*in the same breath*] *chvilku nato* [*a moment later*]
adjectival phrases	*další* [*other*] *jiný* [*different*]
numeral phrases	*první – druhý...* [*the first – the second...*]
verbal phrases	*předcházet* [*to precede*] *následovat* [*to follow*] *zdůvodnit* [*to give reasons*] *způsobit* [*to cause*] *kontrastovat* [*to contrast*]
adverbial phrases	*později* [*later*] *přesněji* [*more precisely*] *původně* [*initially*]
prepositional phrases	*v rozporu s tím* [*in conflict with this*] *kvůli tomu* [*because of that*] *nemluvě o* [*not speaking of*] *na rozdíl od toho* [*unlike that*] *z tohoto důvodu* [*for this reason*] *v důsledku* [*as a consequence*] *v této souvislosti* [*in connection with this*] *pro tento účel* [*for this purpose*]
interjectional phrases	*pravda* [*true*] *tím spíš* [*rather*] *právě tak* [*just as*]
(semi-)clauses	*důvod je jednoduchý* [*the reason is simple*] *výjimkou je* [*the exception is*] *výsledkem je* [*the result is*] *jako příklad uvedl* [*he gave an example*] *stručně řečeno* [*shortly speaking*]

Table 11.2: Czech secondary connectives as syntactic phrases

examples from the PDT in Table 11.2. The aim of this part of our analysis was to find out which syntactic structures Czech secondary connectives belong to in most cases.

The analysis demonstrated that the most numerous Czech secondary connectives are: (i) verbal phrases, (ii) prepositional phrases, (iii) secondary connectives functioning as a (semi-)clause.

11.5.1 Secondary connectives realized by verbal phrases

The most numerous group of secondary connectives (they form 37% of all secondary connectives annotated in the PDT) contains verbal phrases that are both lexically and formally free. They function as secondary connectives in all their paradigms (i.e. they do not function as secondary connectives only in some grammatical forms like *přeloženo* [*translated*] or in collocation with some other words like *jednoduše řečeno* [*simply speaking*] or *jak je vidět* [*as seen*]). For example, the verbs like *zdůvodnit* [*to justify*], *předcházet* [*to precede*] etc. can function as secondary connectives in many variant forms, see Examples 161 and 162:

(161) *Gyula Horn se vyslovil pro možné zavedení majetkové daně.* ***Zdůvodnil*** *to tím, že utahování opasků se nemůže vztahovat pouze na lidi žijící ze mzdy.* (PDT)

Gyula Horn has spoken in defense of the possible introduction of property tax. He ***justified*** *it with the fact that tightening of belts cannot be applied only to people living only off wages.*

(162) *Hranice jedné miliardy Kč by banka chtěla dosáhnout koncem roku 1996.* ***Předcházet bude*** *řada postupných kroků.* (PDT)

The bank would like to reach the limit of one billion CZK by the end of 1996. This ***will be preceded*** *by a series of gradual steps.*

In these examples, the lexical bases of secondary connectives are the verbs *zdůvodnit* [*to justify*] and *předcházet* [*to precede*] signaling with their lexical meaning the semantic type of given discourse relations (i.e. *reason–result* in Example 161 and *asynchrony* in Example 162).

11.5.2 Prepositional phrases

The second most numerous group within the annotated Czech secondary connectives are prepositional phrases (making up 25% of all 1,161 tokens of annotated secondary connectives in the PDT). These expressions consist of two parts – mostly by a secondary preposition, e.g. *kromě* [*in addition to*], *kvůli* [*due to*], *na rozdíl od* [*unlike*], *na základě* [*on the basis of*], *navzdory* [*despite*], *přes* [*in spite of*], *vinou* [*due to*], *vzhledem k* [*considering*] etc. and an anaphoric expression referring to the previous discourse argument, see Example 163:

(163) *Prezident Fernando Collor si údajně nahrabal do vlastní kapsy milióny.* ***Kvůli tomu*** *pravděpodobně padne.* (PDT)

President Fernando Collor probably pocketed millions for himself. ***Because of this****, he will most likely lose power.*

In this example, there is a secondary connective *because of this* signaling a discourse relation of *reason–result* between the first argument (*President Fernando Collor probably pocketed millions...*) and the second argument (*he will most likely lose power*). At the same time, it is here fully replaceable by the primary connective *therefore*. The anaphoric part of the secondary connective (the demonstrative pronoun *this*) coreferentially refers to the previous discourse argument (i.e. *this* semantically contains the whole information that *President Fernando Collor probably hoarded millions to his own pocket*), which is a general feature of this type of secondary connectives.

The lexical core of this type of connectives is formed by the secondary prepositions that also signal given types of discourse relations (e.g. the preposition *because of* signals mostly a discourse relation of *reason–result*, the preposition *in spite of* expresses a relation of *concession* etc.). The secondary preposition is at the same time a fixed part of this type while the anaphoric expressions may vary – we may find several different realizations of these secondary connectives in authentic texts, like *na rozdíl od* ***toho/ této skutečnosti/předchozího****...* [*unlike* ***this/this situation/the previous fact****...*].

11.5.3 Secondary connectives realized by (semi-)clauses

The third most numerous group (with 17% of the total of secondary connectives in the PDT) contains Czech secondary connectives realized by (semi-)clauses (i.e. structures containing either a finite or a non-finite verb).

Most of these secondary connectives contain a finite verb with a weak lexical meaning like *být* [*to be*], *tvořit* [*to form*], *sloužit* [*to serve*], *uvést* [*to give*]. The core of the lexical meaning is carried here by another component (mainly by a nominal phrase) – cf. for example, *důsledkem je* [*the consequence is*], *rozdílem je* [*the difference is*], *výjimku tvoří* [*the exception here is*], *jako příklad slouží* [*to serve as an example*], *jako důvod uvádí* [*to give the reason as*]. This is the reason why we do not classify them under the verbal phrases (where the lexical core lies on the verb, see structures like *to znamená* [*this means*], *to bylo způsobeno* [*this was caused*] etc.).

They are also unique because some of these secondary connectives may be syntactically higher than the second discourse argument (i.e. they may have a form of the governing clauses and the discourse argument is syntactically dependent on them), see Example 164:

(164) *Hráč brazilského týmu napadl v dnešním utkání svého protihráče.* ***To je důvod, proč*** *nebude hrát příští tři zápasy.* (PDT)

The Brazilian football player attacked his opponent in today's match. ***This is the reason why*** *he will not play in the next three matches.*

In this example, there is a secondary connective *this is the reason* in the form of a main clause and the second discourse argument (*he will not play in the next three matches*) is formally a dependent clause.

Other secondary connectives may be realized even by a separate sentence – i.e. some of the multiword discourse expressions have an ability to stay outside the two arguments they connect and to form independent text units (Rysová and Rysová, 2014), see Example 165:

(165) *S vašimi akciemi se musí obchodovat na burze, ale Wall Street vám nabízí cenu z RMS.* ***Důvod je vcelku jednoduchý.*** *V RMS je cena většiny akcií nižší než na burze.* (PDT)

Your shares must be traded on the stock market, but Wall Street offers you a price from the RMS.[122] ***The reason is quite simple.*** *In RMS, the price of most stocks is lower than on the stock market.*

The sentence *the reason is quite simple* in this example expresses a discourse relation of *reason–result* between the two discourse arguments. In this example, it is replaceable by the modified connective *simply because*.

11.6 Lexical Characteristics of Secondary Connectives in Czech

From the lexical point of view, English AltLexes in the PDTB (Prasad, Joshi and Webber, 2010) are examined in terms of their lexical stability, i.e. whether they are lexically free or fixed; each AltLex is classified into one of these two categories. In our project, we do not understand the free and fixed lexical expressions as two closed or separated categories but as a scale with two opposite end points (as in Howarth, 1998 and Howarth, 2000).

The first represents Czech secondary connectives containing a certain key word that enters into several free combinations (both grammatically and lexically unrestricted). Typical examples are verbal phrases (see above) functioning as secondary connectives in the whole paradigm and forming open-ended free collocations, see examples with the verb *dodat* [*to add*][123] found in the Prague Dependency Treebank (in the sense of saying as a further remark): *k tomu je třeba dodat* [*it is necessary to add*], *dodal* [*he added*], *dodejme* [*we should add*] etc.

The second end point of the scale concerns multiword phrases functioning as secondary connectives only in given combinations or forms (like *jak je vidět* [*as seen*] etc.), i.e. the individual lexical items do not function as secondary connectives separately. Thus, for example, the verb *to see* does not signal any discourse relation on its own, cf. *Peter saw his friend*.

Such expressions are lexically and grammatically restricted and are based on certain irregularity (Čermák, 2007). These secondary connectives either exhibit a slight degree of variability, i.e. they occur in a limited set of combinations like *jednoduše/ krátce/obecně řečeno* [*simply/shortly/generally speaking*], or they are fully frozen – occurring only in one possible combination like *tím spíš* [*all the more*].

[122] Czech Stock Exchange

[123] More details on verbs of saying expressing discourse relations in Rysová (2014b).

Secondary connectives standing closer to the second pole of the scale are usually incomplete grammatical structures called *lexical bundles* that are characterized as the most frequently co-occurring sequences of words involved in the organization and structuring of the text (Biber and Conrad, 1999). Most of these secondary connectives have a function of clause modifiers (the so-called disjuncts), i.e. they comment on the content or the style of given clauses, see Example 166 with the multiword expression *jednoduše řečeno* [*simply speaking*] expressing a discourse relation of *generalization*:

(166) *Každý odklad nejenže přináší velké ztráty na dané investici, ale také se nepříznivě promítá do ekonomiky země i veřejného života. Pokud budeme do vysokorychlostní železnice investovat v potřebném optimálním čase, můžeme využít všech jejich výhod. Se zpožděním naopak žádné výhody nezískáme.* ***Jednoduše řečeno***$_{\textbf{generalization}}$*, čím déle budeme projekt odkládat, tím vyšší pak budou náklady.* (PDT)

Every delay not only generates big losses for the given investment, but it also adversely affects the country's economy and public life. If we invest in high-speed rail in the required optimal time, we can take advantage of all the benefits. By delaying, on the contrary, we gain nothing. ***Simply speaking***$_{\textbf{generalization}}$*, the longer we delay the project, the more it will cost.*

As we mentioned above, not all of the secondary connectives may be strictly categorized either as a fully lexically free or fixed expression – see e.g. the secondary connective *sloužit jako příklad* [*to serve as an example*]. This structure is not fully frozen (or idiomatic), as it is not an incomplete grammatical structure, the predicate may be conjugated, the noun *příklad* [*example*] may be modified like *sloužit jako* ***hlavní*** *příklad* [*to serve as a* ***main*** *example*] etc. On the other hand, the structure is neither fully free, as it exhibits a certain degree of expectations and predictability typical for the fixed collocations.

For this reason, we do not apply the strict lexical categorization for the Czech secondary connectives but we conceptualize them as a scale going from the free combinations to idiomatic collocations. In the Prague Dependency Treebank, the majority of secondary connectives occur closer to the free combinations end of the scale.

11.7 Semantic Characteristics of Secondary Connectives in Czech

From the semantic point of view, secondary connectives have a special position within other cohesive means (like reference, substitution or ellipsis) – they signal a discourse relation within a text and at the same time, they contain (implicitly or explicitly) an anaphoric expression referring to the first discourse argument (Forbes-Riley, Webber and Joshi, 2006). In English, the anaphoric reference may occur on the surface (analytical) layer (like *as a result of* ***that***) or not (*as a result*) (Prasad, Joshi and Webber, 2010). The situation in Czech seems to be similar to English. Some of the Czech secondary connectives may optionally express the anaphoric reference in the surface

Reference type	Obligatory	Optional
implicit	*jednoduše řečeno* [*simply speaking*] *přeloženo* [*translated*] *jak je vidět* [*as seen*] *stejným dechem* [*in the same breath*]	*důsledkem je* [*the consequence is*] *výsledkem je* [*the result is*] *důvodem je* [*the reason is*] *příkladem je* [*the example is*]
explicit	*díky tomu* [*thanks to this*] *kvůli tomu* [*because of this*] *i přes tato fakta* [*despite these facts*] *to kontrastuje s* [*this contrasts with*]	*důsledkem tohoto je* [*the consequence of this is*] *výsledkem toho je* [*the result of this is*] *důvodem toho je* [*the reason of this is*] *příkladem toho je* [*the example of this is*]

Table 11.3: Implicit and explicit anaphoric reference (examples)

(like *příkladem* ***toho*** *je* [*the example of* ***this*** *is*] vs. *příkladem je*[124] [*the example is*] etc.), some must express the anaphoric reference in the surface (like *kvůli* ***tomu*** [*because of* ***this***] etc.) and some cannot express the anaphoric reference in the surface (like *stručně řečeno* [*simply speaking*] etc.), see Table 11.3.

The table captures the individual secondary connectives in Czech according to whether the presence of the anaphoric reference is implicit or explicit. Thus, the obligatorily implicit category means that the given secondary connective does not have an ability to express the anaphoric reference in the surface layer. For example, it is impossible to say **k tomu stručně řečeno* [**simply speaking to this*], but only *stručně řečeno* [*simply speaking*].

On the other hand, the obligatorily explicit category contains secondary connectives that would be ungrammatical without the explicit anaphoric reference – we cannot say **Jsem nemocný. Kvůli budu doma.* [**I am ill. Because I will be at home.*], the anaphoric reference is here obligatory, e.g. *Kvůli tomu / této skutečnosti budu doma.* [*Because of this / this fact, I will be at home.*].[125] The category of secondary connectives with optional anaphoric reference means that we have two options – either to express the anaphoric reference explicitly or implicitly.

Whether the secondary connectives express the anaphoric reference in the surface or not is connected closely to their lexico-syntactic nature. The structures that do not have an ability to express the reference in the surface are lexically frozen collocations that are not combinable with other lexical units, i.e. not with an anaphoric reference (cf. *jak je vidět* [*as seen*]).

[124] Generally, Czech does not have articles like English, see Chapter 1. Thus, in the Czech example, there is no explicit indication of reference to the previous context.

[125] For more details, see Rysová and Mírovský, 2014b.

On the other hand, the secondary connectives with obligatorily expressed anaphoric reference are structures that require the presence of another component due to their valency, as in the case of *kontrastovat* [*to contrast*], *znamenat* [*to mean*] or prepositions like *kvůli* [*due to*], *nazdory* [*in spite of*] etc. It is important to say that such expressions function as secondary connectives only in combination with anaphoric reference (referring to the whole previous argument), cf. Examples 167 and 168:

(167) *Nemohu spát* ***kvůli*** *bolení hlavy.* (non-anaphoric, non-connective usage)
I cannot sleep ***due to*** *a headache.*

(168) *Byl jsem nejlepší.* ***Kvůli*** *tomu jsem soutěž vyhrál.* (anaphoric, connective usage)
I was the best. ***Due to*** *this, I won the competition.*

Of all the Czech secondary connectives from the PDT (the 1,161 tokens), 55% express the anaphoric reference optionally, 33% obligatorily and 12% cannot express it in the surface layer at all. This fact supports the idea that in Czech lexically frozen collocations (without the anaphoric reference in the surface) are a minority among secondary connectives.

11.8 Summary

In our analysis, we have introduced further possibilities of expressing discourse relations in Czech. We have described the various language means (besides typical examples of conjunctions, some types of particles etc.) that can signal discourse relations within a text. We have called these expressions secondary connectives (as oppose to grammaticalized, lexically frozen primary connectives) and we have analyzed their lexico-syntactic and semantic nature (in comparison with their counterparts in English).

Secondary connectives (forming 5.4% of all annotated discourse connectives in the PDT) appeared to be very heterogeneous – lexically, we deal with expressions that are fixed collocations (*jak je vidět* [*as seen*] etc.) or open collocations (*to znamená* [*this means*] etc.), syntactically, the secondary connectives may be sentence elements (*kvůli tomu* [*because of this*] etc.), sentence modifiers (*jednoduše řečeno* [*simply speaking*] etc.) or separate discourse units (*důvod je jednoduchý* [*the reason is simple*] etc.); semantically, these language means express explicitly or implicitly an anaphoric reference. The most complex study on secondary connectives in Czech (also with statistics introducing the proportion of the individual semantic relations expressed by primary and secondary connectives in the PDT etc.) is available in Rysová (2015).

In this case study, we tried to demonstrate that the study of secondary connectives has its important place in discourse structuring. Secondary connectives represent the middle part of a scale between primary connectives (i.e. expressions whose primary function is to connect two pieces of text) and non-connectives. In this respect, an analysis of secondary connectives may teach us more about both ends of the scale, i.e. about the possible boundaries between explicitness and implicitness of discourse relations, as well as about the diversity of discourse connectives, which may be useful not only for practical discourse annotations in large corpora, but also generally for a better understanding of discourse.

12

Exploration of Weak Coherence and Coherence Disruptions

As speakers and recipients of discourse, we generally assume that the global meaning of a text can be deduced from the *connections of meanings* of its parts, in other words, from the whole structure. If the connection between certain segments of the text is unclear or none is perceived, we are unable to reconstruct the meaning of the whole. Compare the following sequences of sentences:

(169a) *I don't know what to do anymore. The streets are cold and I have nothing to eat.*

(169b) *Meryl Streep is an American actress. The streets are cold and I have nothing to eat.*

Whereas in 169a it is possible to find the semantic relation between the parts, namely the relation of specification (a chain of troubles in someone's life), in 169b there seems to be no obvious relation between the two sentences, if we do not take into account a particular larger context. We can even see 169b as an erroneous sequence of sentences caused e.g. by an inattentive deletion of the middle part of the discourse during editing.

An intuitive assumption of coherence as an unomissible condition of *textuality* is reflected in various approaches to text structure. General conditions of textuality including the condition of coherence were defined by de Beaugrande and Dressler (1981); in the same vein, Halliday and Hasan (1976) introduce *cohesion* as an important feature of the text. The *Rhetorical Structure Theory* describes discourse as a continuous structure of discourse segments where disruptions cannot occur (Mann and Thompson, 1988). On the other hand, there are approaches which do not make any specific assumptions about the shape of the whole discourse structure. For example, the principles of the discourse annotation in the *Penn Discourse Treebank* work with the label *no relation* (*NoRel*; cf. Prasad et al., 2006; see also Chapter 2) which marks places with no detected discourse relations among those annotated in the PDTB; it can be deduced from the annotation scenario that NoRel in the PDTB can capture a deep coherence disruption.

Having annotated Czech texts in the Prague Dependency Treebank for different types of coherence relations (discourse relations, coreference, bridging anaphora, topic–focus articulation), we can contribute now to the theoretical contradiction. Generally, we assume that the texts in the Prague Dependency Treebank are coherent; what we want to do is to verify this assumption and to explore the nature of places in texts where no type of coherence relations was annotated so far. Do they occur in texts? If so, how is the understanding of the text as a coherent whole ensured? And further: If they do occur, do they mean a mistake has been made in building of the structure of the discourse or are they typical for some places in a text? In this way, we can address the question whether every segment of a text has to be formally or semantically connected with the general body of the text.

12.1 Terminology: Coherence Disruption, Weak Coherence and None of the Annotated Relations (No Relation)

Generally, we can observe that *coherence* in a language is a gradual property: In some places in a text it is strong and clear, being ensured by one or more co-occurrent language means simultaneously. In other places, some kind of coherence relation can be found although it is not as obvious as in the case of strong coherence; in these cases, for full understanding of the text certain recipient's effort in the interpretation is necessary. We call them places with *weak coherence*. The third group represents instances were no coherence relation can be found nor inserted and the global interrelationships are impossible to reconstruct, these are henceforth called *coherence disruptions*.

An obvious way to detect places with weak coherence and coherence disruptions in the data annotated for strong coherence relations is to exclude the instances of annotated strong coherence relations and to analyze the rest. What can be achieved in this first step is a group of instances where none of the relations already annotated in the Prague Dependency Treebank could be observed. In addition to actual coherence disruptions, this group of instances will most probably contain various relations which have not yet been explored in the PDT but which do contribute to the text coherence in a specific way (weak coherence). Henceforth, the whole (negatively defined) category with no discourse-related annotation in the PDT so far is called *no relation*.

12.1.1 Unsignaled relations in the RST Discourse Treebank

Research on assumed weak coherence was carried out by Taboada and Das (2013) who looked for signals of rhetorical (coherence) relations in the texts of the *RST Discourse Treebank*, where, originally, no signals (like discourse connectives) have been annotated. They were thus able to also detect the *unsignaled relations*. The broad list of coherence relations and different types of their signals applied in the research includes a small set of connections without any signaling as a result. Three main groups of unsignaled relations were detected in this way, namely *comment*, *summary* and *-shift*. These relations are defined as follows:

- *Comment*: "In a comment relation, the satellite constitutes a subjective remark on a previous segment of the text. It is not an evaluation or an interpretation. The

comment is usually presented from a perspective that is outside of the elements in focus in the nucleus." (Carlson and Marcu, 2001, p. 49)
- *Summary*: "In a summary-satellite relation, the satellite summarizes the information presented in the nucleus. The emphasis is on the situation presented in the nucleus. The size of the summary (the satellite) is shorter than the size of the nucleus." After reversion of the values "nucleus" and "satellite," the same definition holds for the summary-nucleus relation. (ibidem, p. 68)
- *Topic-shift*: "The relation topic-shift is used to link large textual spans when there is a sharp change in focus going from one segment to the other. The same elements are NOT in focus in the two spans." (ibidem, p. 71)

Although we did not look closely at the RST Discourse Treebank within our analysis, it generally inspired our search for new types of relations which have not yet been annotated in the Prague Dependency Treebank.

12.1.2 Treatment of no relation in the Penn Discourse Treebank

Since the discourse annotation in the Prague Dependency Treebank was inspired by the approach of the Penn Discourse Treebank (cf. Chapter 2), the treatment of so-called *no relation* instances in these corpora is similar.

In the Penn Discourse Treebank, the use of this label signals that none of the annotated relations can be applied between the given two arguments (Prasad et al., 2006). The set of relations that are subject to annotation consists of the following types:

- discourse relations with *explicit discourse connectives* (labelled as *explicit*),
- *implicit discourse relations* – without any connective; the connective can be inserted according to the context (*implicit*),
- *entity-based discourse relations* (*EntRel*), and
- discourse relations with *alternative lexicalizations of connectives* (*AltLex*).[126]

All pairs of adjacent sentences within one paragraph were annotated according to the type of connection observed between them. If none of the types enumerated above was applicable, the *no relation* label was marked explicitly between the sentences. Thus, every border between adjacent sentences in the same paragraph was annotated in a certain way in the Penn Discourse Treebank and the text was represented as a continuous chain of annotated types of relations, with possible disruptions at paragraph borders.

The general result of the annotation in the Penn Discourse Treebank reflects the workflow of the annotation. First, the main focus of attention was on the annotation of the explicit discourse connectives which – as the only group within Penn Discourse

[126] Furthermore, *attribution* of discourse arguments to their authors is annotated in the PDTB. Attribution is captured selectively, as a feature of discourse relations with explicit connectives and their alternative lexicalizations and of implicit discourse relations. It is not marked in connections based on either entity relations or elsewhere.

Treebank discourse annotation – were marked not only between adjacent sentences but also between distant arguments. Later, implicit discourse relations have been annotated. During differentiating and marking implicit discourse relations, new connections have been distinguished, which turned out to be different from explicit as well as implicit relations: entity-based relations, alternative lexicalizations and no relation. The possible co-occurrence of more types of relations between two arguments was not followed; thus, the entity-based relations and alternative lexicalizations have been marked in the group of remaining instances after the annotation of explicit and implicit relations only. In the last step, the instances of no relation were annotated in the remaining cases of adjacent pairs of arguments within the same paragraph.

Under these annotation principles, 254 occurrences of the no relation label in a total of 22,141 possible implicit relations[127] (i.e. 1.15%) were found in the PDTB (Prasad, Webber and Joshi, 2014, p. 926). However, this number includes cases where no relation can be observed between adjacent segments but there can still be a certain relation to a distant segment integrating the given argument into the general structure of the text. The adjacency restriction was relaxed in later versions of PDTB-style annotation principles (cf. *Biomedical Discourse Relation Bank*; Prasad, Webber and Joshi, 2014) where implicit relations may hold also between distant sentences in the same paragraph. According to Prasad, Webber and Joshi (2014, p. 926), this step "has reduced the proportion of potential implicit relations that were marked NoRel… to 0.9% in the BioDRB."

12.1.3 Treatment of no relation in the Prague Dependency Treebank

In the Prague Dependency Treebank, the types of coherence relations that were annotated are slightly different than in the Penn Discourse Treebank and so, naturally, the annotation proceeded in a different fashion. The annotation of coherence relations under discussion includes the following groups:

- discourse relations with *explicit discourse connectives* (labelled as *explicit*; see Chapter 2),
- *coreferential relations* (see Chapter 3),
- relations based on *bridging anaphora* (see Chapter 4),[128] and
- discourse relations with *multiword discourse phrases* (see Chapter 11).[129,130]

[127] During the annotation, the possible implicit relations were divided into implicit relations, alternative lexicalizations, entity-based relations and no relations.

[128] Coreference and bridging anaphora in the Prague Dependency Treebank capture similar phenomena as entity-based relations in the Penn Discourse Treebank, see below.

[129] Our approach to multiword discourse phrases is similar to alternative lexicalizations of discourse connectives in the Penn Discourse Treebank.

[130] Another type of coherence relations is the *topic–focus articulation*. Although it concerns a broader context, it is not reflected in the search for no relation, since it deals with internal structure of single clauses.

The four types of coherence relations mentioned above have been annotated in the whole corpus, i.e. on 49,431 sentences.

In the process of the annotation, the analyses of discourse relations on the one hand and those of coreference and bridging anaphora on the other hand were parallel and independent from each other from the very beginning, being perceived as different aspects (perspectives) of discourse coherence. Both, the annotation of coreference and bridging anaphora in the Prague Dependency Treebank and entity-based relations in the Penn Discourse Treebank capture a very similar phenomenon. In the Prague Dependency Treebank, relations between entities (primarily, nominal groups) are captured, whereas discourse relations based on entities are annotated in the Penn Discourse Treebank, i.e. relations not between nominal groups themselves but between discourse arguments (clauses) containing them. However, unlike entity-based relations in the Penn Discourse Treebank, the annotation of coreference and bridging anaphora is not complementary to other types of coherence relations: A pair of discourse arguments can be connected by a discourse relation alongside with coreference relations. Simultaneously, we carried out annotation of discourse relations expressed by multiword discourse phrases which was independent from the annotation of explicit connectives. Hence, co-occurrence of all the four types of relations (explicit discourse connectives, coreference, bridging relations and multiword discourse phrases) within two discourse arguments is possible (see Example 170); they are not complementary to each other.

(170) *I like working in my office on Sundays. The reason is that, for example, I like singing in the empty corridors of the institute.*
<Arg1: <Arg1: *I*$_1$ *like working in my*$_1$ *office*$_{2\text{bridging}}$ *on Sundays.*>>
The reason is that$_{\text{reason–result}}$, ***for example***$_{\text{instantiation}}$,
<Arg2: <Arg2: *I*$_1$ *like singing in the empty corridors*$_{2\text{bridging}}$ *of the institute.*>>

In Example 170, the two arguments (sentences) are connected with several types of relations:

- the primary discourse connective *for example* bears the meaning of *instantiation* (bold, pink),
- the secondary discourse connective *the reason is* connects the arguments in the *reason–result* relation (bold, green),
- the pronouns *I* and *my* in the first argument are coreferential with the pronoun *I* in the second one (underlined, co-indexed with number 1),
- the nominal group *empty corridors* has a bridging relation to the expression *office* (underlined, co-indexed with number 2).

Another important distinction from the Penn Discourse Treebank is that all four types of coherence relations, namely relations with explicit connectives, relations based on coreference and bridging anaphora, and relations with multiword discourse phrases, have been marked not only between adjacent sentences, but anywhere in a text, disregarding paragraph boundaries. On top of these types of relations, further stylistic and structural aspects of texts and their segments – such as text genre,[131] titles and subtitles, captions of photos, etc. – were marked in the whole extent of the corpus.

Thus, the current stage of coherence annotation in the Prague Dependency Treebank covers the above-mentioned types of expressed signals of coherence and does not directly represent a continuous chain of discourse arguments; rather deep structural relations are being captured. A concept for future research on further aspects of discourse coherence, such as implicit discourse relations and no relation, is being formulated in the first annotation probes. The results of the experimental annotation of *implicit relations*[132] which was performed on 100 sentences are presented in Poláková et al. (2013). The first steps of the research on *no relation* in the Prague Dependency Treebank are the topic of the present chapter.[133]

Data and workflow

We conducted an experiment where no relations were searched for as indicated in Section 12.1. First, the places with no relation were annotated: the label *no relation* was inserted in places in the text where none of the existing annotated relations could have been applied. For the experiment, 38 documents (909 sentences)[134] were chosen proportionally from different text genres.

Since the exploratory annotation should reveal new perspectives of the text analysis in the PDT, we decided to go through all possible connections in texts manually and to judge them individually, even though there was a possibility to exclude the annotated instances of coherence relations automatically.

Instances of no relations were searched for between adjacent arguments only, regardless of paragraph borders. Our initial idea was to provide all slots between two possible neighboring discourse arguments with labels in the following order:[135]

- discourse relations with explicit discourse connectives,
- discourse relations with multiword discourse phrases,
- implicit discourse relations,
- relations based on the coreference and bridging anaphora, and
- no relation.

[131] For the annotation of *text genres* in the Prague Dependency Treebank see Chapter 2; compare also Poláková, Jínová and Mírovský (2014).

[132] The term *implicit discourse relation* stands for a discourse relation corresponding to one of the set of discourse semantic types which is deducible from the context, even if an explicit discourse connective is not used between the arguments.

[133] *Attribution* has not been distinguished in the Prague Dependency Treebank yet.

[134] In this case, a *sentence* is understood as a unit between two final punctuation marks regardless of its internal structure.

[135] The overlap of more types of relations is possible, except for *no relation*.

Originally, the annotation was supposed to progress from expressed forms of discourse connectives (primary discourse connectives and multiword expressions), the annotation of which was the easiest for the annotators, through implicit discourse relations, coreference and bridging relations to the remaining group of no relation, similarly as in the Penn Discourse Treebank.

However, it turned out that such a structure of the work forces annotators to focus more on the difficult issues which are not connected with the research of no relations. While no relations were easy to find intuitively (as a place where neither explicit nor any type of easily inferable connections could be recognized), distinguishing between coreference-based relations and implicit relations proved to be demanding and time consuming. Therefore the sequence of steps in the workflow was modified as follows, in order to proceed from easier tasks to more complex ones:

- discourse relations with explicit discourse connectives,
- discourse relations with multiword discourse phrases,
- no relation,
- relations based on the coreference and bridging anaphora,
- implicit discourse relations,
- (final check)

Parallel occurrence of two or more types of relations between two arguments is possible.

12.2 Results and Discussion

We expected no relation to be a rather rare phenomenon. For this reason, in the first step we analyzed not only the types of connections between regular discourse arguments (i.e. clauses with finite verbs, cf. Chapter 2), but also isolated nominal groups and other types of semi-clauses if they were delimited as sentences (by final punctuation marks or by paragraph indent, e.g. in titles of articles like *Possible solutions*; from now on, we will call them *incomplete discourse arguments*). Hence, in this type of annotation, every text was represented as one continuous chain of marked connections including, among others, incomplete discourse arguments, as well as paragraph borders. The annotation includes intra-sentential and inter-sentential discourse relations. The results are described in Table 12.1.

In order to keep the research consistent and comparable with previous types of discourse annotations in the Prague Dependency Treebank (which did not concern incomplete discourse arguments), we have applied the second measurement concerning relations between full discourse arguments only. In this case, the texts are not represented as continuous chains of relations. Although connections crossing paragraph borders are still annotated, connections relating incomplete discourse arguments are not captured anymore. The results are presented in Table 12.2.

All types of relations (number of instances)	1,176
No relation (number of instances)	95
No relation within all types of relations (%)	8.8

Table 12.1: Proportion of *no relation* in annotation including incomplete discourse arguments

All types of relations (number of instances)	1,107
No relation (number of instances)	27
No relation within all types of relations (%)	2.44

Table 12.2: Proportion of *no relation* in annotation limited to complete discourse arguments

As follows from the comparison of Tables 12.1 and 12.2, the presence of incomplete discourse arguments decreases[136] the number of recognized coherence relations significantly: full arguments are connected more clearly than incomplete discourse arguments (2.44% and 8.8% of no relations, respectively). A closer look at the data shows that incomplete discourse arguments have often specific discourse functions: they are text titles and subtitles, names of cities where the news was published, text organizing remarks such as the caption *photo* relating to an image in the newspaper, they present the names of the authors of the article, their acronyms, etc. These structures, henceforth called *text-organizing devices*, are typical for the journalistic texts taken as the language material for the Prague Dependency Treebank; high number of their occurrences in the data is therefore not surprising. The occurrences of no relation including relations with incomplete arguments were further classified into two large groups:

- Arguments connected by a new type of relation, not yet captured in the annotation (approx. 75% of all occurrences of no relation). In this group, relations based on text-organizing devices are absolutely predominant (ca. 50% of all occurrences of no relation), followed by attribution (ca. 25% of all occurrences of no relation). The relation of question–answer is represented marginally.
- Sentences with assumed no relation. It turned out that comprehensibility in these cases is based on the reader's expectation of a coherent text (approx. 25% of all occurrences of no relation). After taking into consideration the perspective of the reader's expectation, no occurrence representing a real coherence disruption remained in our data.

[136] Or more exactly, fewer relations from the list given above can be detected between incomplete discourse arguments and other parts of the texts.

It is important to remember that the numbers describing the proportions of individual groups are only approximate, since in some cases more than one relation between two arguments can contribute to the coherence of the discourse (e.g. the relation between a title and the text itself together with a list structure of the text).

12.2.1 New types of relations

According to the methodology of annotation of no relation described in Section 12.1, we expect to detect new types of coherence relations which were not present in the originally annotated coherence relations. In this section, we focus on three types of such coherence relations, namely on *text-organizing devices*, *attribution* and the relation between *question and answer*. These relations form well distinguishable groups with typical semantic and/or formal features; in the terminology of Taboada and Das (2013), they are often signaled explicitly, however, by other language means than annotated in the previous stages of the coherence annotation in the Prague Dependency Treebank.

Text-organizing devices

The text-organizing devices do not have to be connected with the adjacent text by any of the means listed in 12.1. Still, they do contribute to text coherence and they are integrated into the text. They organize and segment the text by giving some meta-textual information, e.g. the theme of the text, its author, the source of the information, etc., cf. Example 171.

(171) *Po několika letech usilovného jednání ruská strana ústy svého premiéra Viktora Černomyrdina vyjádřila ochotu splatit dluhy vůči České republice.* [**subtitle**]
[original annotation: NoRel][137]
Miroslav Svoboda [**author**]
[original annotation: NoRel]
Tyto dluhy nevznikly najednou. [**basic text**] (PDT)

After several years of intensive negotiations the Russian side, through its Prime Minister Viktor Tchernomyrdin, expressed its willingness to pay back its debts to the Czech Republic. [**subtitle**]
[original annotation: NoRel]
Miroslav Svoboda [**author**]
[original annotation: NoRel]
These debts did not arise all at once. [**basic text**]

[137] Two phases of analysis are captured in the example sentences. The first of them, called the original annotation of no relations, shows between which segments no relation was found originally. The second one marks a new interpretation of the relation (bold), e.g. relation between the title and the text; eventually, parts of sentences connected with the relation are highlighted (underlined), if necessary.

The relations based on text-organizing devices connect other types of text segments than discourse relations or coreference and bridging anaphora do. Whereas the latter connect a certain argument with a similar argument in the text, the text organizing relation holds between a sentence (or more sentences, or an incomplete discourse argument) and the text as a whole, the basic body of the text. In the proper linear reading of the text, a text-organizing device is recognized and understood as a meta-text commentary. As such, it is excluded by the reader from building a chain of relations between adjacent arguments. It has a certain relation to the neighboring sentences but this relation is not direct, it is mediated via the whole of the text. So, in Example 171, *Miroslav Svoboda* is not the author of the immediately preceding and following sentences only, but of the whole text.[138]

Some types of text-organizing devices can be recognized automatically: first, they are often placed at typical positions within the text; second, they can be marked by specific graphics in a written text. Furthermore, they may be expressed by typical lexical expressions, e.g. the name of an author is a named entity, a name of a person. (Typically, it is expressed in an isolated sentence.)

Attribution

Another type of relation can be observed between a discourse segment containing information and a discourse segment containing the source of the information. Both notions are understood very broadly in this analysis: The term *information* may mean direct or indirect speech, a thought, a statement, generally the communicated content; the term *source of information* covers the author of the information but also e.g. the document containing the information (cf. *the Bible says*). In the following example, the *source of information* and *information* are underlined:

(172) *Našemu listu se podařilo získat od představitelů slovenského Úřadu pro normalizaci, metrologii a zkušebnictví (ÚNMS SR) informaci o technickém zajištění propouštění potravinářských výrobků do SR. Z rozhodujících opatření, která by měla plně vstoupit v platnost po 1. dubnu, vyjímáme:*

[original annotation: NoRel][**attribution**]

1) Do 31. března platí v plném rozsahu postup podle dohody ÚNMZ ČR a ÚNMS SR, na jejímž základě český výrobce (slovenský dovozce) získá na základě schválení české zkušebny a rozhodnutí Ministerstva zdravotnictví SR na ÚNMS SR potvrzení o platnosti rozhodnutí i na území SR. (PDT)

Our newspaper managed to obtain information about technical details of how food products are admitted into the Slovak Republic from representatives of the Slovak Office of Standards, Metrology and Testing (ÚNMS SR). We extract the following

[138] The text-organizing devices differ from discourse arguments in one more regard: The feature of "being text-organizing device" is characteristics of the unit itself, not of the relation.

from the decisive arrangements which are supposed to be fully operational after April 1st:
[original annotation: NoRel][**attribution**]
(1) Until 31 March, the procedure runs according to the agreement between ÚNMZ ČR [Czech Office for Standards, Metrology and Testing] and ÚNMS SR [Slovak Office of Standards, Metrology and Testing] to a full extent. According to the agreement, a Czech producer (Slovak importer) will obtain for his approval by the Czech testing office and the decision of the Ministry of Health of the Slovak Republic a confirmation of validity of the decision for the region of the SR from ÚNMS SR.

The units of this relation are often not discourse arguments on both sides: a thought as a discourse argument is often ascribed to an entity expressed by a nominal group (author, document). In fact, *attribution* may be understood as one of bridging relations connecting parts of the text in a specific way.

From the formal point of view, a probable presence of attribution can be expressed by certain signs, such as verbs of saying and thinking, punctuation (colon, quotation marks), personal names or lexical items denoting texts which can be quoted (*the law, the announcement* etc.).

Question and answer

Another relation which has not been marked in our annotation but which is sure to establish a coherence relation is the connection between *question and answer*,[139] cf. Example 173.

(173) *Bude vláda postupovat podle svých původních záměrů?*
[original annotation: NoRel][**question–answer**]
Ano. (PDT)

Is the government going to follow its original plans?
[original annotation: NoRel][**question–answer**]
Yes.

This type of relation is introduced in other theories of text structure, cf. Taboada and Das (2013, p. 254) referring to the *solutionhood relation* between question and answer. In particular, questions are signaled with typical signs such as question intonation, reverse word order, interrogative words and question mark. Although answers can be marked by specific language means, too, (particles *yes/no*, elliptical sentences, introducing particles like *well*, etc.), they do not need to be signaled at all. Therefore, in some cases a pair of adjacent question and an immediately following argument can be misunderstood as a structure related as question and answer, although in fact, the

[139] Beside this relation, questions and answers can be connected by other types of relations, too, e.g. with coreference (talking about the same entity).

second unit is not connected to the previous question in the above sense, cf. rhetorical questions or Examples 174 and 175. In these examples, the questions are parts of the title or the subtitle whereas the sentences that follow belong to another part of the text (subtitle and author's name, respectively); thus, the relation of question and answer is not established here.

(174) *Co by vám stálo za další utažení opasku?* [title][**+question**]
Anketa čtenářů [subtitle][**–answer**] (PDT)

What would be worth tightening your belt once more? [title][**+question**]
Readers' survey [subtitle][**–answer**]

(175) *Československé manažerské centrum připravuje druhý ročník soutěže o cenu Vynikající podnikatel.* [title]
Loni prvenství získal Petr Chodura, zakladatel firmy Chodura z Ostravy. [subtitle]
Kdo bude úspěšný letos? [subtitle][**+question**]
Jan Hábík [author][**–answer**] (PDT)

The Czechoslovak manager center is preparing the second annual competition for Remarkable Entrepreneur prize. [title]
Last year, Petr Chodura, the founder of Ostrava firm Chodura, won the title. [subtitle]
Who will have success this year? [subtitle][**+question**]
Jan Hábík [author][**–answer**]

In case of questions and answers, it will be an interesting task to automatically detect the possible placement and extent of the second argument – answer. This is a specific field of research which has not yet been addressed in detail in the Prague Dependency Treebank.

12.2.2 Reader's expectation as coherence factor

Having excluded the new types of relations (see Section 12.2.1) from the detected set of occurrences of no relations, we arrive at a general question about the very notion of *text* illustrated in the introduction of this chapter: How do we discern a pair of random sentences from a text consisting of two sentences if the coherence is not signaled? Why do we accept a TV program as one text whereas all the inscriptions along one street are not considered to represent a single text?

The cue is the *reader's expectation of coherence* (cf. Hobbs, 1979; Kehler, 2002), his or her experience with text structures applied to a given chain of sentences. If we find the first signs of *textuality*,[140] we want to validate it, to find further features of textuality, the global meaning of the text being one of them. Therefore, we are ready to insert

[140] First signs of possible textuality are e. g. common graphics of all parts, co-occurrence on the same page, etc.

an assumed global meaning into the sequence of sentences; we are well disposed to "excuse" the lack of explicit expressions of coherence and to choose the most coherent interpretation of the connection from all the possibilities.

Based on our experience, we expect a certain set of textual features of the text. In this section, the following formal as well as semantic features are described which help to establish text coherence: *thematic continuity* together with *world knowledge*, and *genre rules*. The expectation of coherence influences the *reconstruction of ellipses* and *cataphora understanding*, too.

Expectation of thematic continuity and coherence based on world knowledge

The expectation of coherence works in local connections as well as on a higher level, between larger units where the reader is ready to look for a thematic continuity and to understand it as a sign of text coherence. The *thematic continuity* is deduced from cue words which are accepted as a part of a net of related notions; obviously, the readers' ability to recognize such a net of notions in a text is dependent on their world knowledge. If the world knowledge is insufficient, the necessary inference is impossible, and the pair of text spans is incoherent for the recipient.

In a similar vein, relations between items of a *list* can be interpreted if the general net of notions is recognizable for the reader. This net consists of the notion of hyperonymy (generality of one object over other objects) and a set of relations between the *hyperonym* and the *hyponym*. This net allows the reader to understand the relations between even heterogeneous items as *cohyponymic relation*.[141]

The expectation of a certain structure triggered by a signal of the list is a principle on which coherence understanding is based in some text genres. Following the thought that items in a list are connected in some way and that they build a coherent unified structure, we are able to accept e.g. an overview of a TV program as a coherent text rather than as isolated fragments. Similarly, larger structures like specific text genres can be based on the list structuring principle and be accepted as coherent, if related to a common hyperonym, cf. articles bringing mutually unrelated news from a certain field like *World news*, *Last week*, *Qualification matches for European Championship*. This structure is illustrated in Example 176 where news from the world of computers are listed:

(176) *Týden mezi počítači* [title]
[News describing a lesson on geographic informational systems, with the following final sentence:] *Seminář byl součástí mezinárodní konference Evropa v pohybu: kontext GIS.*
[original annotation: NoRel]
[Another piece of news:] *1. září letošního roku zahájila činnost nově vzniklá soukromá škola – Škola výpočetní techniky s. r. o.* [Description of the school follows.] (PDT)

[141] The individual items of a list can reach a large extent, e.g. of whole paragraphs.

The week among computers [title]
[News describing a lesson on geographic informational systems, with the following final sentence:] *The lesson was held as part of the international conference Europe on the Move: Context GIS.*
[original annotation: NoRel]
[Another piece of news:] *On 1 September of this year, a newly established private school Computing School, Ltd. began its operation.* [Description of the school follows.]

In Example 176, the original *no relation* annotated in our corpus can be later reinterpreted as a cohyponymic relation of isolated news from the world of computers; the general topic, indicated in the title, works as a hyperonym connecting the seemingly disparate paragraphs. However, without the overarching notion, the coherence line would be lost and the sense of connecting these paragraphs into one document would be unclear.

Genre rules

The example with lists is connected to the most general type of coherence covering the text as a whole – the expectation of meeting *genre rules*. Having recognized the genre of a text, we expect not only its thematic unity, but we also make assumptions about its whole structure which helps us especially in genres with a certain degree of a structural standardization. During reading or listening, the incoming parts of a text are inserted into the supposed structure according to which they are interpreted. In this way, the supposed structure works as a main coherence frame even if the signals of coherence are not explicitly expressed between the segments themselves. This can happen e.g. in an interview where the reporter does not develop the answers of the interviewee in further questions, but he or she sticks to a prepared list of questions disregarding the flow of the actual dialogue. The assumed structure of an interview still helps the reader to bridge the emerging disruptions. Similarly, overview genres (e.g. schematic descriptions of matches and their results in sports news, financial information) are based on the knowledge of the respective genre rules – of the typical information they provide and the form in which the information is presented. See Example 177 bringing standardized weather information (it was shortened, only the main parts of the structure are presented here):

(177) *Ozón: Koncentrace ozónu v ozónové vrstvě nad naším územím je 9 % pod dlouhodobým průměrem.*
[original annotation: NoRel]
Slunce: Erupční aktivita slabá, geomagnetické pole slabě porušené.
[original annotation: NoRel]
Předpověď na dnešek: …
[original annotation: NoRel]

Sobota a neděle: …
[original annotation: NoRel]
Ozón: …
[original annotation: NoRel]
Slunce: …
[original annotation: NoRel]
Rekordy dne: Nejvyšší teplota 33,6 st C byla v roce 1828, nejnižší 8,5 st C v roce 1948. Dlouhodobý teplotní normál – 19,5 st C.
[original annotation: NoRel]
Přímořská letoviska … (PDT)

Ozone: Ozone concentration in the ozone layer over our territory is 9% below the long-term average.
[original annotation: NoRel]
The Sun: The eruption activity is weak, the geomagnetic field is slightly disrupted.
[original annotation: NoRel]
Forecast for today: ...
[original annotation: NoRel]
Saturday and Sunday: ...
[original annotation: NoRel]
Ozone: ...
[original annotation: NoRel]
The Sun: ...
[original annotation: NoRel]
Records of the day: The highest temperature of 33.6 °C was in 1828, the lowest of 8.5 °C in 1948. The long-term average temperature is 19.5 °C.
[original annotation: NoRel]
Seaside resorts …

The overview in Example 177 presents an enumeration of criteria and their values. Unlike in pure lists presented in the previous subsection, this structure is more complicated: The criteria are grouped according to days (the weather for today, for Sunday and Monday etc.), there is a certain hierarchy between the items of the list. The reconstruction of the coherent whole is the reader's task and it is supposed to be based on his or her knowledge of the genre.

Reconstruction of ellipses

Coherence is closely related to *ellipses* and their reconstruction. Generally, omission of a segment is possible if its reconstruction is simple, i.e. in cases where the coherence stays clear disregarding the missing parts. This results in the fact that the absence of explicit signals of coherence may signify either an incoherent text or, on the other hand, a strongly coherent text where the signals of coherence could have been omitted. In our data, this concerns especially nominal groups whose parts related to the previous context are elided, see Examples 178 and 179.

(178) *Relativně tak stát vynakládá na tržně konformní podporu malého a středního podnikámí přibližně 1,6–1,8 % hrubého domácího produktu.*
[original annotation: NoRel]
Regionální aspekt [*podpory*, **elided**][subtitle][142] (PDT)

Thus, the state spends about 1.6–1.8% of the gross domestic product, in relative terms, on the market-based support of small and medium enterprises.
[original annotation: NoRel]
Regional aspect [*of the support*, **elided**][subtitle]

(179) *Do 31. května 1994 by měla být podepsána dohoda, která by se měla zabývat restrukturalizací ruského dluhu. Zároveň by měla být impulzem pro uzavření dalších dlouhodobých dohod o dodávkách plynu a ropy do ČR.*
[original annotation: NoRel]
Různá řešení [*dluhu*, **elided**][subtitle] (PDT)

Until 31 May 1994 an agreement should be signed dealing with the restructuring of the Russian debt. At the same time it should be the impetus for the conclusion of further long-term agreements on gas and oil supplies to the Czech Republic.
[original annotation: NoRel]
Different solutions [*of the debt*, **elided**][subtitle]

In both examples, the elided parts specify nouns with general meanings; since definiteness is not obligatorily expressed in Czech (see Chapter 1), no signals of coherence are given explicitly in the nominal groups. We suppose that it is the general meaning of the governing nouns which is unclear in isolation and thus encourages the reader to reconstruct the ellipsis, cf. *the regional aspect of what? different solutions of what?* We further assume that similar phenomena can be expected in structures with elided subjects, which are very common in Czech (cf. Chapter 1).[143]

[142] As in Example 174, the interpretation of coherence relations is supported by the relation of the subtitle to the basic body of the text. Due to the rare occurrence of no relations in a text we have not found a pair of arguments where the coherence is based on the reconstruction of an ellipsis only. Nevertheless, there is no doubt that appropriate reconstruction of ellipses is essential for understanding text relations.

[143] The unambiguousness of reconstructed elements differs. Whereas a subject can be undoubtedly reconstructed, ellipses like *different solutions of something* can be reconstructed in more ways (*solutions of the debt, of the problem, of the question*). Nevertheless, there is still a coreference or bridging relation between the reconstructed expression and some part of the previous text.

Distant cataphoric connection

Reconstruction of ellipses described in the previous part is mostly based on a search for a proper antecedent. Generally, most coreferring expressions refer anaphorically, not cataphorically. Whereas a search for an antecedent connects a new piece of information to the old one, search for a postcedent in *cataphoric relations* connects a new piece of information coming first to another one, expected to come in the following text. In this way, temporary incoherence can occur, especially in the cases of distant cataphoric connections. In most cases, the first part of the relation cannot be understood before the second part occurs: The unclear status of the first part is thus reinterpreted and newly incorporated into the general idea of the text after the introduction of the second part. This way of posterior clarification of the relations in the text (and a possible point where temporary *no relations* can occur) is typical e.g. for titles and subtitles whose relation to the forthcoming text gets clearer only at the end of the segment.

12.3 Summary

The analysis of occurrences of weak coherence and coherence disruptions in our data presents two observations. First, three systematic groups of coherence relations in text were identified which contribute to the common interconnection of the text segments and which have not been annotated yet in the Prague Dependency Treebank, namely *text-organizing devices*, *attribution* and the relation *question–answer*.

Although the remaining instances of no relation did not constitute a consistent category, a unifying principle could be found which explains the relation of these segments to the rest of the text. The identification of the relation is based on the *readers' expectation of a coherent text* and on their willingness to search for an appropriate coherent interpretation of a seemingly incoherent sequence of sentences. The expectation of coherence works in a local environment as well as globally (search for coherence based on *thematic unity* or on *genre rules*, pro-coherent interpretations of *ellipses* and *cataphoras*).

Coming back to the question set in the introduction of this chapter whether all parts of a text have to be mutually interconnected, we can now state that our data do not contain a counterexample. We did not detect any type of coherence disruption whose connection to the rest of the text could not be later uncovered. As far as the data allowed, we can say that the texts are coherent with no exceptions. Nevertheless, to find the continuity, i.e. the coherence signals, a large apparatus of complex mechanisms is needed, such as reinterpretation and insertion. Regarding future annotation directions, the analysis and classification of the no relation group revealed the existence and nature of further discourse structuring mechanisms.

13

Contextually Bound Expressions without a Coreference Link

In Chapters 3, 4 and 5, we introduced the Prague approach to the topic–focus articulation, coreference and bridging relations, we have also presented how these phenomena are captured in the Prague Dependency Treebank. Let us now look at how these three phenomena interact. In this chapter, we will explore interdependencies of contextual boundness of nominal groups and all types of coreference, anaphoric and bridging relations which are registered in the PDT.

As described in Chapter 5, an expression is considered to be contextually bound if the speaker can assume that it refers to an object easily identifiable by the recipient, i.e. if he considers it to be uniquely determined by the context. It is natural then to suppose that contextually bound nouns and nominal groups are linked to their antecedents by some kind of anaphoric relations,[144] such as grammatical coreference, textual coreference, bridging relation, or a reference to a text segment that have been annotated in the PDT (see the detailed descriptions in Chapters 3 and 4).

To illustrate our expectations, we provide Examples 180–182.[145] In Example 180, the contextually bound expression *oba kluky* [*both guys*] refers back to *Steve Wozniak* and *Steve Jobs*. The expression *přepážka ve Spálené ulici v Praze* [*the desk in Spálená street office in Prague*] in Example 181 is connected by a bridging relation of a set–subset type with *Česká pojišťovna* [*the Czech Insurance company*]. In Example 182, the contextually bound expression *tento hlas* [*this opinion*] refers to the whole preceding segment of direct speech.

Coreference:

(180) *V Cupertinu se sešli dva známí z dětských let:* ***šestadvacetiletý Steve Wozniak*** *a* ***o pět let mladší Steve Jobs****.* ***Oba kluky***$_{\mathrm{t}}$ *spojovalo bezmezné nadšení pro počítače a touha vyrobit stroj svých snů – opravdový osobní počítač.* (PDT)

In Cupertino, two friends from childhood met: ***twenty-six-year old Steve Wozniak*** *and* ***a five-years younger Steve Jobs****.* ***Both guys***$_{\mathrm{t}}$ *were full of enthusiasm for computers and desire to create a machine of their dreams – a true personal computer.*

[144] In this chapter, the word *anaphoric* is also used to refer to coreferential and bridging links together.

[145] Here and in further examples, non-contrastive contextually bound items are labelled with *t*, contextually non-bound nodes are marked as *f* and contrastive contextually bound items are marked with *c*.

Bridging relations:

(181) *Avšak v případě pojištění nabízeného* ***Českou pojišťovnou*** *je důležitou skutečností, že pojištěný má mimo sjednanou pojistnou částku zaručenou zvláštní prémii a navíc valorizaci... O prémiích jsme se bohužel* ***u přepážky***$_{\mathrm{t}}$[146] ***ve Spálené ulici v Praze*** *nedozvěděli, úřednice se... zmínila pouze o valorizaci.* (PDT)

However, in case of insurance offered by ***the Czech Insurance company****, it is important to know that besides the agreed sum, an insured person has an assured special premium and valorization... As for premiums, unfortunately, we were not informed about them* ***at the desk***$_{\mathrm{t}}$ ***in Spálená street office in Prague****, the clerk... mentioned only the valorization.*

Reference to a text segment longer than one sentence:

(182) *„Ne, já jsem tu petici proti těm dětem nepodepsal. Oni je vyhnali ještě dříve, než se sem nastěhovali. Nikdy jsem ve vztahu k člověku neslyšel větší cynismus než termín sociální skládka.“* ***Tento hlas***$_{\mathrm{t}}$ ***staršího důchodce*** *je v Košťanech zcela ojedinělý.* (PDT)

"No, I did not sign the petition against those children. They drove them out before they moved here. I have never heard a more cynical thing said about people than the term social landfill." ***This viewpoint***$_{\mathrm{t}}$ ***of an older pensioner*** *is in Košťany quite unique.*

A cursory analysis of typical examples and literature on this topic (see e.g. Daneš, 1979)[147] supports our expectations. But is it always the case, that contextually bound expressions are linked by an anaphoric link? Can there be some systematic reasons for contextually bound nouns without any coreference or bridging anaphoric links? Can the reasons be classified? What is the nature of these reasons – are they rather technical or they have a deeper theoretical background? What will we find out about the text coherence and its representation in the PDT if we analyze anaphoric links and contextual boundness together and what novel information does the classification of contextually bound nouns without anaphoric links reveal?

[146] In this chapter, we mark the contextual boundness only with the expressions in question. Their dependent nodes may have different values of this attribute (see Chapter 5 for more detailed description of this issue).

[147] In Daneš (1979), the author also mentions other reasons for contextual boundness, such as knowledge coming from life experience and situational knowledge. However, by analyzing the reasons for contextual boundness, he takes into account only the cases with antecedents explicitly expressed in the previous context.

Contextual boundness and anaphoric links	Number of nodes
contextually bound nodes without anaphoric links	21,529 (30%)
contextually bound nodes with an anaphoric link:	
with textual coreference links	37,606 (54%)
with bridging links	6,755 (10%)
with grammatical coreference links	3,709 (5%)
with reference to text segment	876 (1%)
all contextually bound nodes	69,583 (100%)

Table 13.1: General statistics of contextually bound nodes with/without anaphoric links

13.1 Data

To answer these questions, we collected the PDT statistics of non-contrastive contextually bound semantic nouns (tectogrammatical attribute *tfa*=*t*).[148] For these nodes, we then considered in how many cases they are linked by grammatical/textual coreference, bridging relations, or reference to a text segment. Contextually bound nodes that do not have any kind of anaphoric reference (grammatical coreference, textual coreference, bridging relation, or reference to a text segment) form a special class. The statistics for nominal groups that are explicitly expressed in the sentence is presented in Table 13.1.

As we can see from Table 13.1, almost one-third of contextually bound expressions in the PDT are linked neither by coreference nor by associative bridging relations. To find out the reasons why it is so and how "the context" still "binds" these expressions, we randomly selected 500 of these cases from the PDT texts and analyzed their boundness from the formal, grammatical, semantic and pragmatic points of view.

In our analysis, we only considered elements that are present at the surface level. The statistics for newly established nodes in the tectogrammatical structure (reconstructed nodes in case of ellipsis) is somewhat different. According to the definition, newly established nodes are mostly understood from the context and, as such, should be marked as contextually bound. Contextually non-bound new nodes and contrastive contextually bound nodes (making together ca. 4% of all elided nominal expressions) are limited to (i) list structure root nodes representing identification structures (titles), (ii) foreign-language expressions where the value of contextual

[148] We selected the following types of nominal expressions: core nouns and possessive adjectives (tectogrammatical attribute *sempos*=n.denot), deverbal nouns ending with *-ní* / *-tí* such as *plavání* [*swimming*] and deadjectival nouns ending with *-ost* such as *nezralost* [*immaturity*] (*sempos*=n.denot.neg), demonstrative pronouns in the positions of syntactic nouns (*sempos* = n.pron.def.demon) and personal pronouns and their possessive counterparts (*sempos* = n.pron.def.pers).

boundness is assigned to the foreign-language expression as a whole and (iii) textual ellipsis of the governing noun, as in the case of Example 183 (Figure 13.1), where the omitted noun *záležitost* [*affair*] for *krátkodobá záležitost* [*short-term affair*] is reconstructed in the tectogrammatical structure and, as a first mention, it is a contextually non-bound node (labelled as *f* in the dependency tree).

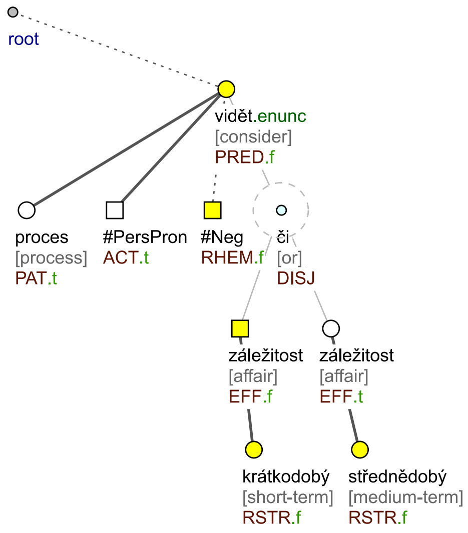


Figure 13.1: Newly established nodes: contextually non-bound nominal groups

(183) *Proces nevidí jako krátkodobou* [***záležitost***]$_{\mathrm{f}}$ *či střednědobou záležitost*$_{\mathrm{t}}$. (PDT)
He does not consider the process to be a short-term [***affair***]$_{\mathrm{f}}$ *or a medium-term affair*$_{\mathrm{t}}$.

Within contextually bound reconstructed nominal groups that are scrutinized in our analysis, almost 90% (15,552 instances) are linked by anaphoric relations. An analysis of cases without anaphoric relations has shown that the reasons for such absence are similar to those concerning expressions present at the surface level. However, for newly established nodes, there are more errors in human coreference annotation.[149]

13.2 Reasons of Contextual Boundness without Anaphoric Links

The analysis of 500 first cases of contextually bound nominal expressions without coreference and bridging reference has revealed that reasons for this situation can be divided into the following groups:

[149] In some cases, coreference and bridging relations should have been marked, but have been overlooked by human annotators due to the complexity of the tectogrammatical structure of the given sentence.

Group of reasons	Number of nodes
contextual boundness is deduced from some kind of semantic or pragmatic relation to the previous context	157 (32%)
a nominal group refers to secondary circumstances (temporal, local, etc. modifications)	110 (22%)
contextual boundness of a nominal group has extralinguistic reasons	30 (6%)
contextually bound expressions represent rates, degrees, scales, proportions, etc.	102 (20%)
technical reasons	88 (18%)
annotation errors	13 (2%)
total number of analyzed occurrences in the PDT	500 (100%)

Table 13.2: Explanation of the absence of anaphoric links by contextually bound expressions

- contextual boundness is deduced from some kind of semantic or pragmatic relation to the previous context;
- contextual boundness of nominal groups has extralinguistic reasons (expressions referring to unique objects in the given situation etc.) or is given by common world knowledge;
- nominal group expresses mainly secondary circumstances (e.g. temporal and local settings);
- contextually bound expressions represent rates, degrees, scales, proportions, etc.;
- contextually bound expressions are not linked by any anaphoric link for technical reasons.[150]

The distinction between these five groups is not exact. Moreover, the lack of anaphoric links with a contextually bound expression in analyzed data may be explained by more than one reason with the same instance as mentioned above. Thus, the statistics of reasons for contextual boundness without anaphoric links presented in Table 13.2 is very approximate. However, we believe it will help the reader understand the relations between the different types.

In Sections 13.2.1 through 13.2.5, we will analyze each group separately and provide examples from the PDT 3.0.

[150] There is also a minor group of contextually bound nominal groups which are not linked by coreference or bridging relations by an error, or they are mistakenly marked as contextually bound by the annotators.

13.2.1 Deduction of contextual boundness from previous context

In our sample, the most frequent cases (32%) were those where contextual boundness (marked by *tfa* attribute *t*) is deduced from some kind of semantic or pragmatic relation to the previous context close to bridging relations annotated in the PDT, but not annotated in such a way, since an explicit specification and classification of such cases is beyond the current understanding of bridging relations.[151]

For example, the relation between *pojišťovny* [*insurance companies*] and *své produkty* [*their products*] in Example 184 could be interpreted as a bridging set–subset relation. Indeed, the notion of insurance company is closely semantically related to the products they offer. On the other hand, this relation is not straightforward and does not unambiguously refer to a set–subset type.

(184) ***Pojišťovny****, které povolení Ministerstva financí nemají, u nás* ***své produkty***$_{\mathrm{t}}$ *podle zmíněného zákona nabízet nesmějí.* (PDT)

According to this law, ***the insurance companies****, which do not have authorization from the Ministry of Finance, are not allowed to sell* ***their products***$_{\mathrm{t}}$ *here.*

Another example (Example 185) represents an associative relation between a natural phenomenon and a person professionally studying this phenomenon (*počasí* [*weather*] and *meteorologové* [*meteorologists*]). Although semantically close, this relation has not been specified as a bridging relation in the annotation of the PDT.[152]

(185) *Přesto se zdá, že největší nadějí na zmírnění vlny* [*uprchlíků z Kuby do USA*] *je* ***bouřlivé počasí****, které* ***meteorologové***$_{\mathrm{t}}$ *čekají ode dneška.* (PDT)

Yet it seems that the best hope for alleviating the waves [*of refugees from Cuba to the USA*] *is* ***the stormy weather*** *that* ***meteorologists***$_{\mathrm{t}}$ *expect for today.*

The nature of the relation between entities in Example 186 is more ambiguous.

(186) *Ještě stále méně nákladné jsou platby za dodávky dotovaného tepla než investice do* ***zlepšení izolačních vlastností objektů*** *a do* ***dalších opatření***$_{\mathrm{t}}$ *ke zlepšení tepelné pohody v nich.* (PDT)

Paying for subsidized heat is still less costly than investing in ***improvements of insulating properties of the buildings*** *and* ***other enhancements***$_{\mathrm{t}}$ *to improve the level of thermal comfort in them.*

The relation between *zlepšení izolačních vlastností objektů* [*improvements of insulating properties of the buildings*] and *dalších opatření* [*other enhancements*] in Example 186 is

[151] For a more detailed description of bridging relations annotated in the PDT, see Chapter 4.

[152] The reasons for not annotating such relations is the chosen approach to annotate only six specific groups of bridging relations, leaving other types of bridging inferences unattended. The justification for this decision is supplied in Chapter 4.

close to what we marked as a bridging relation of type *ANAPH* (explicit anaphoric relation without coreference, see Chapter 4) or bridging relation of type *CONTRAST*. However, these relations have not been marked by annotators. As for *CONTRAST*, the relation *x* – *other x* (represented as *improvements* – *other enhancements* in 186) does not fit the definition of contrast of nominal groups given in Chapter 4. As for *ANAPH*, the absence of a bridging relation of this type has rather technical reasons: the adjective *další* [*other*] has not been characterized as a possible anaphoric adjective in the guidelines for annotation of coreference and bridging relations in the PDT (Nedoluzhko and Mírovský, 2011). Indeed, strictly speaking, *other* is not an anaphoric adjective in a restricted sense, it is something in between contrastive and anaphoric, and therefore, it is not marked as a bridging relation.

Contextual boundness of nominal expressions can be given by anaphoric adjectives, such as *další* [*another*] and *pomocný* [*additional*] in Example 187. These relations have co-hyponymic nature in a narrow or broader sense, which are not annotated as bridging in the tectogrammatical structure in the PDT.[153]

(187) ***Určitým signálem pro posouzení kvality cestovní kanceláře*** *je prospekt a dokumentace, kterou vás vybaví na začátku jednání...* ***Dalším hlediskem***$_{\mathrm{t}}$ ***při vašem rozhodování*** *by mělo být hlavní teritoriální zaměření kanceláře...* ***Pomocným kritériem***$_{\mathrm{t}}$ ***při volbě*** *je také chování personálu při jednání o koupi zájezdu.* (PDT)

A special signal to evaluate a travel agency's quality *is documentation and booklets, which will be provided for you at the beginning of the talk...* ***Another factor***$_{\mathrm{t}}$ ***in your decision*** *should be the territorial focus of the travel agency...* ***An additional test***$_{\mathrm{t}}$ ***for your choice*** *is how the staff behaves toward you during the conversation about purchasing a vacation.*

Another typical case of contextual boundness which is deduced from associative relations is a variation of semantically connected topics related to the same so-called *general topic* (*hypertopic*) of the text. For example, the sentence 188, is extracted from the text consisting of more than 40 sentences that informs the reader about how difficult it is to get a grant and how much time and energy scientific organizations spend on preparing grant proposals. In this context, the noun *boj* [*fight*] is unambiguously contextually bound, although it was not explicitly mentioned in the preceding text.

(188) ***Boj***$_{\mathrm{t}}$ *o získání grantů se tak stává novou profesí, která je daleko více svého druhu uměním nežli vědou.* (PDT)

Thus, ***the fight***$_{\mathrm{t}}$[154] *for obtaining grants is becoming a new profession that is much more similar to some kind of art than to science.*

[153] The reasons for not annotating co-hyponymy are discussed in detail in Chapter 4.

[154] Definite article is absent in Czech (see Chapter 1).

High frequency of contextually bound expressions without an anaphoric link is also due to the nature of journalistic texts in the PDT, which contain many formal characteristics, such as addresses (including streets and building numbers), lists with sports results, news and so on. In Example 189, the expression *výhra* [*win*] is repeated six times, but these are all different wins, so they cannot be interpreted as coreferential.[155] In a very broad sense, the relation between such nominal groups could have co-hyponymic nature, but as we have already stated, this relation is not captured in the PDT either.

(189) *V prvním tahu 18. týdne Sportky v I. pořadí není* ***žádná výhra****, ve II. jsou 3* ***výhry***$_{\mathrm{t}}$ *po 166237 Kč, ve III. je 89* ***výher***$_{\mathrm{t}}$ *po 8005 Kč, ve IV. je 4973* ***výher***$_{\mathrm{t}}$ *po 286 Kč, v V. je 86407* ***výher***$_{\mathrm{t}}$ *po 35 Kč.* (PDT)

In the first draw of Sportka's 18th week, there was ***no win*** *in the first sequence; in the second, there are 3* ***wins***$_{\mathrm{t}}$ *worth 166,237 CZK; in the third, there are 89* ***wins***$_{\mathrm{t}}$ *worth CZK 8,005; in the fourth, there are 4,973* ***wins***$_{\mathrm{t}}$ *worth CZK 286; in the fifth, there are 86,407* ***wins***$_{\mathrm{t}}$ *worth 35 CZK.*

To summarize, this group consists of the following subtypes:

- contextual boundness is deduced from previous context from relations similar to bridging relations (types *set–subset*, *CONTRAST*, *ANAPH*, etc.); however, these occurrences have not been annotated in the PDT because of their vague nature and high ambiguity in the given context;
- contextual boundness is deduced from the relation of co-hyponyms in a broad sense;
- contextual boundness is deduced from general topic of the text;
- contextual boundness is a part of formal characteristics (addresses, sport lists, etc.); such cases are frequent and the contextually bound expression may repeat several times, increasing the number of expressions in this group.

13.2.2 Extralinguistic reasons for contextual boundness

This group consists of contextually bound expressions without an anaphoric link, the contextual boundness of which cannot be deduced from the preceding context and has extralinguistic reasons. These are, for example, references to objects that are unique in the given situation (Example 190), deictic references without a deictic element (Example 191), a reference to the common knowledge of the speaker and addressee (Example 192), etc. The information can also be understood as "given" (and marked as contextually bound in the PDT) if the addressee can derive it based on his world knowledge together with logical inferences drawn from the previous context (Example 193).

[155] The expression *výhra* [*win*] in Example 189 may also be considered as a rate, as described in Section 13.2.4.

(190) *Kompletní informace pro drobného investora v* ***LN***$_{\text{t}}$ [*Lidových novinách*] *na dvou stránkách.* (PDT)

Complete information for small investors in ***LN***$_{\text{t}}$ [*Lidové noviny*] *is on two pages.*

Here, *LN* is an abbreviation for *Lidové noviny*, one of the most popular newspapers in the Czech Republic, from which the text of this article is excerpted for the PDT. A reader is holding this newspaper in his hands when reading this sentence, thus its name is contextually bound by the situation itself and does not need any special introduction.

(191) ***Dvoustranu***$_{\text{t}}$ *připravil Jaromír Složil.* (PDT)

The[156] ***two-page spread***$_{\text{t}}$ *was prepared by Jaromír Složil.*

The situation in Example 191 is similar to Example 190. The difference here is that *dvojstrana* [(*the*) *two-page spread*] is not a named entity, it has even more extralinguistic reference and could be used with a demonstrative pronoun in the same context without any substantial change in meaning.

In the following Example 192, contextual boundness of the nominal group *návrh příslušných smluv* [*the proposal of the relevant agreements*] is deduced based on the world knowledge of the author and addressee: Such a big political exchange presupposes signing a high number of contracts.

(192) *O výměně Bojnického oltáře za deset gotických deskových obrazů slovenské provenience se dohodli zástupci ministerstev kultury ČR a Slovenské republiky. Podle tiskové mluvčí českého ministerstva kultury Evy Rolečkové* ***návrh***$_{\text{t}}$ ***příslušných smluv*** *předloží slovenská strana do 15. května.* (PDT)

The representatives of the Ministries of Culture of Czech and Slovak Republics agreed on the exchange of Bojnický altar for ten Gothic panel paintings of Slovak provenance. According to the spokeswoman of the Czech Ministry of Culture Eva Rolečková, the Slovak side will submit ***the proposal***$_{\text{t}}$ ***of the relevant agreements*** *by 15 May.*

Cases where contextual boundness of an expression is given jointly by the previous context and common world knowledge are close to associative relations of cohesive nature where contextual boundness of an expression is based on the information given in the previous context (the cases described in Section 13.2.1). The borderline between context and world knowledge is not sharp. In order for the addressee to activate the possibility of *soudní spor* [*litigation*] in Example 193, he should know that working for two sports clubs simultaneously may be problematic.

[156] Definite article is absent in Czech (see Chapter 1).

(193) *Zdeno Cíger podepsal s Trenčínem předběžnou roční smlouvu, která mu v případě vyřešení jeho sporů s Oilers umožňuje okamžitý návrat za moře. Zároveň však má platný kontrakt i v Edmontonu. Když NHL nemusí respektovat platné smlouvy hráčů s našimi kluby, tak proč by si totéž nemohl dovolit slovenský hokej?* [...] *V tuto chvíli se **soudního sporu**$_t$ bát nemusí, neboť vztahy NHL a Slovenského svazu ledního hokeje nemají žádný právní rámec.* (PDT)

Zdeno Cíger signed a preliminary one-year contract with Trenčín that allows him an immediate return overseas in case of solving his disputes with the Oilers. At the same time, however, he has a valid contract in Edmonton. If the NHL does not respect existing players' contracts with our clubs, why can Slovak clubs not do the same? [...] *At the moment, he need not worry about possible **litigation**$_t$, because there is no legal framework for the relations between the NHL and Slovak Ice Hockey Association.*

13.2.3 "Scene setting" circumstances

This group comprises of adverbial modifiers with the meaning of "scene setting" circumstances, mostly temporal and local ones. Their function in discourse is to orient the addressee in time and space, to position the speaker's statement so that it would be properly understood. Nominal expressions in such modifications are in topic and are contextually bound, but they do not need to be previously mentioned: Their contextual boundness is evident from the situation of speech.

The temporal description may be very general (Example 194) or more specific (Example 195):

(194) *U posudků **v minulosti**$_t$ mohl být sebemenší náznak negativního hodnocení spouštěcím mechanismem pro šikanování.* (PDT)

***In the past**$_t$, the slightest hint of a negative rating in the review could cause bullying.*

(195) ***V prosinci**$_t$ **minulého roku** vzniklo v Hudebním divadle v Berlíně detašované pracoviště budoucího institutu.* (PDT)

***In December**$_t$ **of last year**, an off-site working space of the future institute was founded in the Musical Theatre in Berlin.*

In Example 196, the local circumstance *v některých státech* [*in some countries*] makes the statement about structured cabling by reconstruction of administrative buildings less general, claiming that it is true only in some countries.

(196) *Dalším trendem, **v některých státech**$_t$ při výstavbě nebo rekonstrukci zejména administrativních budov dokonce předepsaným, je strukturovaná kabeláž.* (PDT)

*Another trend, which is even prescribed for the construction or reconstruction of administration buildings **in some countries**$_t$, is structured cabling.*

Temporal and local secondary circumstances may be present together within a single modification in the sentence. For example, in 197, the adverbial modifier *na včerejší tiskové konferenci* [*at yesterday's press conference*] refers both to when and where the statement was uttered.

(197) *Na* ***včerejší tiskové konferenci***$_{\mathrm{t}}$ *to řekla zástupkyně Zeleného kruhu Marie Haisová.* (PDT)

Marie Haisová, the deputy of Green Circle, said it at ***yesterday's press conference***$_{\mathrm{t}}$.

The modifications may also specify the statement with regard to some characteristics, as shown in Example 198.

(198) ***Z hlediska chování***$_{\mathrm{t}}$ *je tato třída dobrá.*

As for behaviour$_{\mathrm{t}}$, *this class is a good one.*

Within the Firbasian framework of the theory of FSP and his metaphoric view of functional sentence perspective as a theatrical scene (Firbas, 1992), temporal, local and other circumstances of this type are associated with a coulisse. This was disputed in Sgall, Hajičová and Buráňová (1980), claiming that such descriptions can also be in focus, with a high degree of communicative dynamism and contextually non-bound. See, e.g., the modification of Example 195 in 199, with another word order in Czech, where the temporal specification *v prosinci minulého roku* [*in December of last year*] is in focus position and contextually non-bound.

(199) *Detašované pracoviště budoucího institutu vzniklo v Hudebním divadle v Berlíně* ***v prosinci***$_{\mathrm{t}}$ ***minulého roku***. (PDT)

The off-site working space of the future institute was founded in the Musical Theatre in Berlin ***in December***$_{\mathrm{t}}$ ***of last year***.

13.2.4 Contextually bound expressions representing measures

Contextually bound nominal groups have no anaphoric reference to the previous context if they represent different kinds of measures (rates, degrees, scales, proportions, etc.) that are standard for measuring the given items. Contextual boundness of such items can be explained by their low referential potential. When serving as measures, nominal expressions function rather as parameters of measurement, since their reference to objects as such in this case is moved to the background. Thus, not being referring in proper sense, they are much less probable to take part in coreferential relations.

In Example 200, the amount of wheat (3,718 million) is contextually non-bound and new, and the measure (*tons*) is contextually bound.[157]

(200) *Největší objem produkce byl podle Českého statistického úřadu dosažen u pšenice, které se podle odhadů sklidilo 3718 mil.* ***tun***$_{\mathrm{t}}$*.* (PDT)

According to the Czech Statistical Office, the largest amount of production was achieved in wheat, which was estimated to 3,718 million ***tons***$_{\mathrm{t}}$*.*

Other nominal expressions may also serve as measure, e.g. *lidé* [*people*] in Example 201.

(201) *Během téže doby zaměstnanost na letišti vzrostla na 2200* ***lidí***$_{\mathrm{t}}$*.* (PDT)

During the same period, employment at the airport increased to 2,200 ***people***$_{\mathrm{t}}$*.*

If measure is expressed in terms of time, some cases are close to examples in the previous Section 13.2.3 (temporal circumstances), see Example 202.

(202) ***Za rok***$_{\mathrm{t}}$ *úřady odebraly jen 4 koncesní listiny.* (PDT)

In the course of the year$_{\mathrm{t}}$*, the authorities revoked only 4 concession documents.*

The difference from temporal circumstances is clearer when the noun representing temporal specification is modified: The modification will always be contextually non-bound, see Example 203:

(203) *Podle představitelů Slovenské národní jednoty Češi v SNP* [*Slovenské národní povstání*] *obrali Slováky o vlastní stát a* ***na 40 let***$_{\mathrm{t}}$ *jim vnutili čechokomunistický režim.* (PDT)

According to representatives of the Slovak National Unity, Czech people in SNP [*Slovak National Uprising*] *stole from Slovaks their state and imposed the Czech-communist regime on them for* ***40 years***$_{\mathrm{t}}$*.*

In this case, the time period has a relatively high degree of communicative dynamism and the number of years (40) is contextually non-bound. Only the measure noun itself, in this case *let* [*years*] is contextually bound.

[157] Contextual boundness of measures in Examples 200 and 201 may also be explained by world knowledge: To a Czech reader, it is generally known that wheat is measured in tons and that employment concerns people. However, in the manual annotation of contextual boundness, this is one of the problematic points – annotators often mark measures as contextually bound although they are neither clear from the previous context nor directly based on the world knowledge. We believe that reasons of frequent annotators' errors in such cases is the low referential potential of nouns in this function.

13.2.5 Technical reasons

This group includes cases where contextually bound nodes have certain anaphoric reference to the previous context, but there is no coreferential arrow leading from this node in the annotation of textual coreference and bridging relations in the PDT. However, in most cases, the anaphoric relation can be easily reconstructed for this node according to a set of heuristic rules.

A contextually bound nominal expression may miss an anaphoric link when it is a part of a larger nominal expression which already has an anaphoric link. Such cases can often be explained by syntactic structures of dependency trees on the tectogrammatical layer. It concerns, for example, coordinative and oppositional constructions with reconstructed elided nodes.

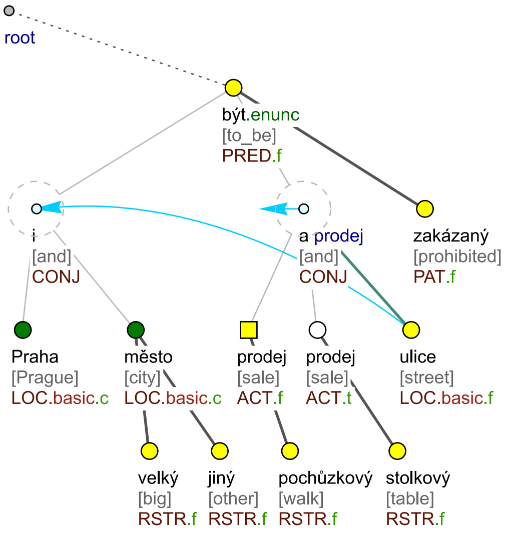


Figure 13.2: Contextually bound expression as a dependent phrase of a larger nominal group with a bridging relation (Example 204)

(204) *V Praze i v jiných velkých městech je* ***pochůzkový*** **[*prodej*]** ***a stolkový prodej***$_t$ *na ulicích zakázaný.* (PDT)

In Prague and other big cities, ***walking*** **[*sale*]** ***and table sale***$_t$ *on the streets is prohibited.*

This is also the case for first names that are dependent on second names in dependency trees. Generally with named entities, the absence of anaphoric links by contextually bound nodes is quite common when they consist of more than one nominal expression (see e.g. *financí* [*Finance*] within *Ministerstvo financí*$_{t}$ [*Ministry of Finance*] in Example 205).

(205) ***Ministerstvo financí***$_{t}$ *uděluje podle zákona o pojišťovnictví licence pojišťovnám a zároveň vykonává dozor nad jejich podnikáním.* (PDT)

*According to the Insurance law, **the Ministry of Finance**$_{t}$ grants a license to insurance companies and also supervises their business.*

According to the annotation guidelines for the annotation of bridging relations in the PDT, we did not mark the bridging relations in straightforward dependencies, which are marked by some tectogrammatical attributes, e.g. *APP* (*Appurtenance*) and *PAT* (*Patient*).[158] This is the case of Example 206, where the belonging relation is expressed by the dependency relation represented by the tree edge in the tectogrammatical structure of the sentence (see Figure 13.3).

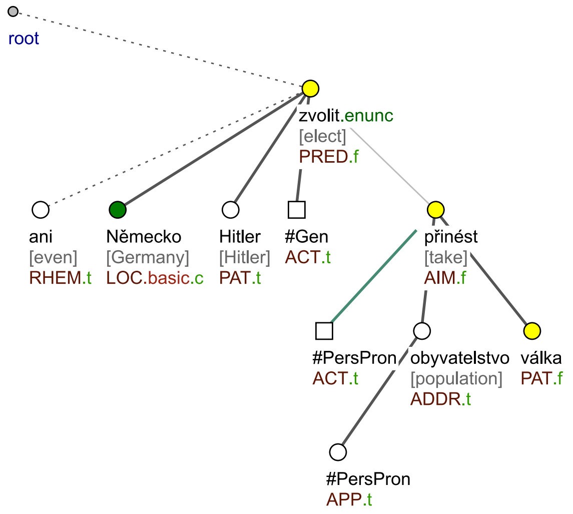


Figure 13.3: Direct dependency with the *APP* functor

(206) *Ani v **Německu** nebyl Hitler zvolen proto, aby jeho **obyvatelstvu**$_{t}$ přinesl válku.* (PDT)

*Even in **Germany**, Hitler was not elected to take his **population**$_{t}$ to war.*

[158] The relations between tectogrammatical attributes and the annotation of bridging relation in the PDT is discussed in Chapter 4, Section 4.4.

Another interesting type of contextually bound expressions without anaphoric links is the case of more or less functional nouns (with lower referential potential) which govern nominal groups with anaphoric links in tectogrammatical trees and together make an anaphoric entity (e.g. *funkce* [*function*] as a governing node for *funkce režiséra* [*the director's function*] in Example 207 and Figure 13.4):

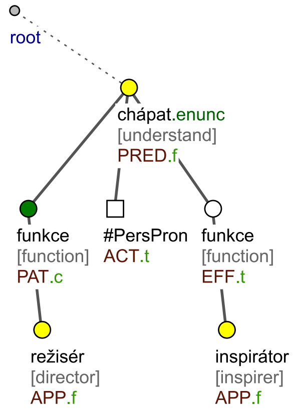


Figure 13.4: Non-referential noun as a governing node (Example 207)

(207) ***Funkci***$_c$ ***režiséra*** *chápu jako funkci inspirátora.* (PDT)
I understand ***the director's function***$_c$ *as that of inspirer.*

Generally, we can summarize that contextually bound expressions in this group lack anaphoric links mostly due to (i) conventions based on tectogrammatical structure of the PDT trees (reconstruction of ellipses, expressions with some tectogrammatical functors, and in (ii) multiword expressions (including construction with functional words). In both cases, anaphoricity of unlinked expressions may be deduced: In the first case, using the tectogrammatical rules, in the second case, with the help of multiword expression processing performed in Bejček, Straňák and Pecina (2013).

13.3 Discussion

We have noticed that the largest group of contextually bound elements without anaphoric links represent relations that can be deduced from different kinds of semantic or pragmatic relationships to the previous context. The semantic relations could be in some cases interpreted as bridging relations but have not been specified as such in the PDT. The reasons for not annotating such cases as bridging were rather diverse. It is clear that all cohesive texts are linked by different kinds of associative relations, but it is not possible to register them all in manual (and even less in automatic) annotation.

A few groups of bridging relations may be singled out more or less precisely (e.g. the part–whole relation annotated in the PDT) but even these relations appear to be vague in real texts. Annotating all possible kinds of associative relations would make inter-annotator agreement extremely low, thus such annotation could never be used for any automatic experiments. More than that, even for scientific goals such information will scarcely be very useful: Texts where everything is somehow connected to everything through relations that have no precise rules will not bring objective positive results. Also, the co-hyponymic relations have not been included into the annotation of bridging in the PDT 3.0. The reasons were mainly pragmatic: In real corpus texts, co-hyponymy tangles with meronymy and set–subset relations so strongly that taking these relations into account makes it almost impossible to make such annotation consistently.

Another finding we have made is that there is a very indistinct border between context and world knowledge. We have tried to classify the cases where contextual boundness was based on the knowledge of the preceding context, or it was derived from it (Section 13.2.1) from the cases based on the world knowledge of extra-linguistic factors (Section 13.2.2).[159] Indeed, there is a significant number of cases that unambiguously belong to one of the defined groups. However, there are many contextually bound expressions that need both contextual and world knowledge. Moreover, as for contextual interpretation, it can also be based on understanding the interconnections between the preceding elements and the contextually bound expression in question. This is the case in Example 193 described above and Example 208 below.

(208) *Zastřelený lesník. Kladruby. Lesník s prostřelenou hlavou byl nalezen v Kladrubech na Tachovsku. Dvacetiletý muž pracoval u Lesní společnosti Stříbro. Policisté prokázali, že z* ***legálně držené kulovnice***$_{\text{t}}$ *vypálil osudnou ránu sám.* (PDT)

Forester shot dead. Kladruby. A forester shot through the head has been found in Kladruby in the Tachov region. A twenty-year old man worked for the Forestry company Stříbro. The police proved that using ***a legally held shotgun***$_{\text{t}}$*, he made the fatal shot himself.*

In Example 208, the meaning of the expression *legálně držená kulovnice* [*legally held shotgun*] can be activated by the immediately preceding expressions *lesník* [*forester*], *zastřelený* [*shot dead*] and *prostřelená hlava* [*shot through the head*]. But is the relation between a forester and legally held shotgun lexicalized enough? Such expressions would be interlinked in WordNet-like databases as a relation *profession–basic instrument*, but is this sufficient to consider such relations to be semantic and thus refer to language alone? In cohesion-based approaches (see e.g. Halliday and Hasan, 1976, and a corpus study based on their conception in Lapshinova-Koltunski and Kunz, 2014), they are considered to be contextual. However, looking at a larger number

[159] The complexity of the distinction between inter- and extra-textual relations was shown e.g. in Kehler and Rohde (2013) on the example of world knowledge influence on pronoun interpretation.

of texts shows that world knowledge is also needed to interpret such cases correctly. Moreover, there are different kinds of professions, some of them are marginal and it would not be easy to link them up to a typical instrument.

Another relevant consideration concerns the degree of referentiality of nominal expressions. When speaking about anaphoricity, we think prototypically about referential nominal groups with specific reference, such as *oba kluky* [*both guys*] in Example 180. However, the further we move from specific expressions with concrete meaning towards nominal groups with predicative meaning (e.g., deverbatives and some abstract nouns), the more complex the question of their ability to take part in coreference relations becomes. The ability of nouns with different referential potential was discussed in Chapter 3. Here, we would like only to point out that it is obviously logical that contextually bound nominal groups with a low referential potential (for example, rates and degrees exemplified in Section 13.2.4) are less probable to have anaphoric links to preceding context because their referential properties are closer to properties of non-nominal non-referring expressions (adjectives, verbs, etc.).

As for technical reasons presented in Section 13.2.5, almost all of them have ”reconstructable” anaphoric links. Thus, although the number of contextually bound nodes in this group is relatively high, it can actually be neglected.

13.4 Summary

In this case study, we analyzed contextually bound nominal expressions explicitly expressed in the sentence, that lack an anaphoric (bridging, coreference or segment) link to a previous context. The statistics collected from the PDT annotated data has shown that in almost one third of contextually bound expressions there is no anaphoric link to the previous context. To analyze the reasons for this finding, we selected 500 random cases and analyzed them in more detail.

Disregarding different types of more or less technical reasons evoked by the tectogrammatical structure of the PDT sentences and annotation errors, we can claim that there are three groups of reasons why contextually bound nominal groups are not linked by any anaphoric link. These are (i) contextually bound nominal groups semantically or pragmatically related to previous textual or extralinguistic context but not specified as bridging relations within the PDT; (ii) secondary circumstances (temporal, local, etc.) and (iii) nominal groups with low referential potential.

14

Tracing Salience

14.1 Motivation

Let's now return to the sample text taken from Josef Škvorecký's book *Dvorak in Love* and focus on the girl child Magda.

> (1) *Across the river* ***Magda*** *and Kovarik could now see a fire with two figures beside it.* (2) *When* ***they*** *moved closer,* (3) ***they*** *could make out two white horses against the background of the dark bushes…* (11) *He looked at* ***Magda****.* (12) *The* ***child's*** *eyes, wide in amazement, stared across the river at this fairy-tale banquet...* (26) *From downstream* ***they*** *could hear a banjo playing. …* (Škvorecký, 1986)

We meet Magda in the very first sentence and continue reading about her in the following two sentences (2) and (3). Then the attention turns to other objects and Magda enters the scene again in sentence (11). We find out even more information about her in the following sentence (12). But then she disappears and comes back again in the sentence (26) and later.

Tracing the appearances and the disappearances of Magda is like hiking in the mountains. Either we are on top of a hill if Magda is in the scene, or we are at the foot of a hill if Magda is in the background. We go up and down depending on whether Magda appears in the scene or not. Instead of kilometers, we measure the distance that we have gone in sentences that we read. We can also draw the route of our walk with Magda, see Figure 14.1.

But Magda's (dis)appearance on the scene is not exactly what one thinks of when reading the text. Although Magda is not present in the scene in sentence (4), she is still on the reader's mind. But as the reader keeps reading, Magda is more and more at the back of his mind. However, she gets back to the forefront of his mind in sentence (11). Therefore, returning to the mountains with Magda, we do not face such steep downhill climbs as during the last trip but we have to climb up the same way, see Figure 14.1.

Our contribution to the discussion of different discourse-related aspects concerns a study of activation of objects in readers' minds while reading the text. So far, we have been intuitively using the terms *one's mind* and *activation* and we placed them in the context of the discourse. Josef Škvorecký shares with the reader a story that is about various objects (characters), like *Magda*, *Kovarik*, *the beauty*, *the black man in*

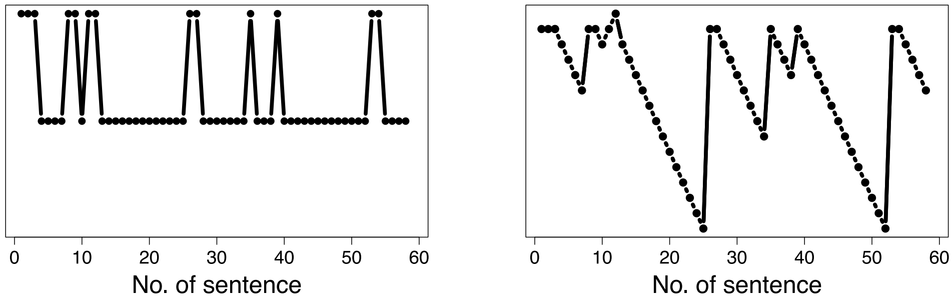


Figure 14.1: Tracing Magda's appearance in the sample text (on the left) and her activation in the reader's mind when reading the sample text (on the right). The sentences flow in the horizontal direction from left to right and they are visualized as points so that the number of points in the plot corresponds to the number of sentences in the text. Appearance and a degree of activation can be read in the vertical direction. The higher the position of a point, the higher is its activation in the reader's mind when reading a given sentence.

livery, *the Master* etc. In other words, the writer's and the reader's minds are kept on the same objects, i.e. on the same knowledge. Formally, we work with a *stock of shared knowledge* shared by the author and the reader or the speaker and the hearer. The *salience* (activation) of each element is changing as the discourse flows and we quantify it using a *degree of salience*, see Figure 14.2 where we illustrate this for six elements from Škvorecký's story, namely for *Magda* [1], *Kovarik* [2], *two figures* [3], *fire* [4], *two horses* [5], *bushes* [6]. We distinguish the elements' salience by the various shades of gray, the darker the color of the element, the more salient the element is, e.g. the child Magda [1] is at the bottom of the stock of shared knowledge before the 10th sentence, then she moves to the very top of the stock in the 11th sentence.

The present chapter focuses on tracing salience in the Prague Dependency Treebank 3.0 (PDT) and is divided into the sections as follows: after a brief overview of related research in Section 14.2, we quickly describe in Section 14.3 the linguistic phenomena annotated in the PDT that present the starting points for the notion of salience that we use. All these phenomena are described in greater details in the previous chapters. Therefore we provide readers only with sample examples simply as a reminder. In Section 14.4, we describe the knowledge-based algorithm according to which a salience degree is assigned to the elements in the stock of shared knowledge. This algorithm was designed at the time where no deeply annotated corpus was available. Thus the PDT offers a great opportunity for verification of the algorithm on a data set containing a significant amount of sentences; so we trace salience in the

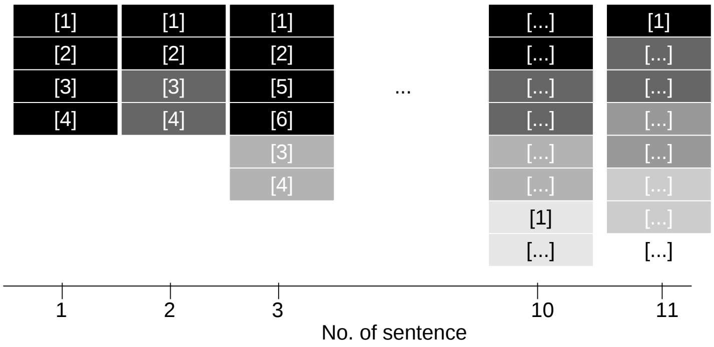


Figure 14.2: Elements in the stock of shared knowledge change their salience.

PDT using the salience algorithm and including visualization of its results. Further we model salience using machine learning techniques and we present our pilot study in Section 14.5.

14.2 Related Work

There are several approaches to discourse dynamics analysis with respect to a sentence structure. Mainly, they attempt to capture the impact of sentence-level expressions on the flow of discourse and its topics. Most of these theories are based on distinguishing two main semantic types of information in the sentence: given vs. new (although their terminology varies, often without significant differences in the definitions).

A three-level hierarchy of givenness of information (contrasting given vs. new) between the speaker and the hearer is proposed by Prince (1981). Each level refers to a different understanding of givenness in previous works:

- givenness as a predictability/recoverability, as defined by Kuno (1972) and Halliday (1967), although their definitions slightly differ,
- givenness in the sense of salience, relating to the assumption of the hearer's consciousness, referring to Chafe (1976),
- givenness in relation to a state of "shared knowledge" according to Haviland and Clark (1974), focusing on what the hearer "already knows and accepts to be true" vs. what the hearer "does not yet know."

Prince then continues with a definition of a more fine-grained familiarity scale for discourse entities, working also with the hearer's ability to infer or link the newly mentioned entities. Another "givenness hierarchy" is presented by Gundel, Hedberg and Zacharski (1993) focusing on the success of nominal expression referents.

Characteristics	*train-1*	*train-all*
number of documents	316	2,533
number of sentences	4,700	38,727
average number of sentences per document	14.9	15.3
number of tectogrammatical nodes	68,626	567,258
average number of tectogrammatical nodes per sentence	14.6	14.6
average number of tectogrammatical nodes per document	217.2	223.9
number of coreference chains	4,519	39,415
average number of coreference chains per document	14.3	15.8
average coreference chain length	3.26	3.25
number of grammatical coreference links	2,226	18,156
number of textual coreference links	7,514	67,535
number of bridging anaphora links	1,987	23,512

Table 14.1: Statistics on PDT training datasets

Another well-known approach to modeling discourse dynamics in terms of sentence structure is the centering theory introduced in Joshi and Weinstein (1981) and further refined by Grosz, Weinstein and Joshi (1995), based on the local attentional states of the speaker and the hearer. It operates with forward and backward looking centers of sentences and defines four types of sentence transitions based on the relations of their centers. One of the typical features of this theory is ranking the centers according to a language-specific parameterization.

An entity-grid model is proposed in Barzilay and Lapata (2008), where each entity occurring in the text (based on coreference relations) is assigned a column in a grid, and each sentence corresponds to a row in this grid. The cells are then filled with syntactic roles of the entities in the corresponding sentence, recording also the transitions between those sentences. It should be noted that this approach, of all the ones we have already mentioned, is the most computationally oriented. Distributional information about the entities are extracted naturally from the entity-grid as well, forming the parameter of salience as a discourse prominence.

An even more application-oriented approach is presented in Sauper, Haghighi and Barzilay (2010), building a statistical-based model of content structure for using it in discourse analysis.

Our approach directly follows the notion of *salience* first mentioned and described in Hajičová and Vrbová (1982), revisited in Hajičová (2003) and further refined and tested in Hajičová, Hladká and Kučová (2006). This notion studies dynamicity of discourse together with the topic–focus articulation of its individual sentences. In contrast to the works mentioned above, this approach postulates a continuous scale

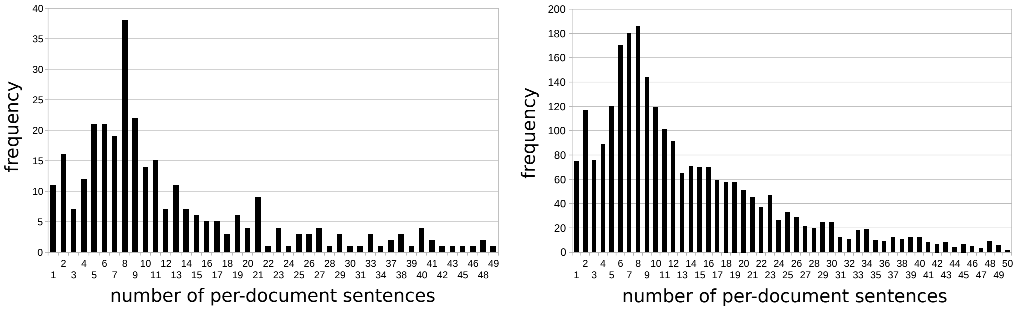


Figure 14.3: Distribution of the number of sentences per document in *train-1* (on the left) and *train-all* (on the right). The cut off is 50 per-document sentences. For illustration, there are 38 documents in *train-1* and 185 documents in *train-all* containing 8 sentences.

of salience and does not suggest any labels for the particular salience levels. Although this might be confusing from the linguistic point of view, it does not have to present difficulty with a computational or machine learning approach.

14.3 Related Linguistic Phenomena annotated in the PDT

The approach to salience that we follow is based on the phenomena of coreference and topic–focus articulation and on their annotation in the PDT.

Prague Dependency Treebank

The Prague Dependency Treebank 3.0 with its 49,431 manually annotated sentences from Czech newspapers is an immense resource for linguistic research in the area of natural language processing, as well as discourse analysis (for a more detailed description of the PDT, see Chapter 6). Technically, the PDT is split into separate training and test subsets, specifically eight training sets *train-[1-8]*, all together called *train-all*. We use the *train-1* and *train-all* data sets only to present numerical data that we can see interesting for modeling salience, see Table 14.1. More detailed distribution of the total number of sentences per document is shown in Figure 14.3. Note that the most frequent number of sentences per document is 8 for both *train-1* and *train-all*, while the average number is 14.9 and 15.8, respectively.

Coreference

Coreference is a linguistic phenomenon describing the relation among two (or more) expressions referring to the same discourse entity in the text. There are backward

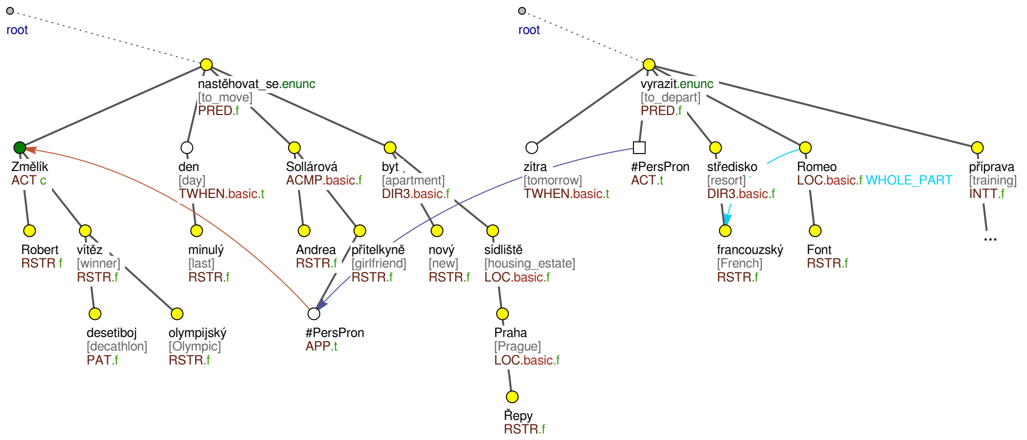


Figure 14.4: Coreference, bridging and contextual boundness annotation on top of the tectogrammatical trees for Example 212

and forward directions recognized in coreference relations with respect to the order of the referring expressions in the text, mainly the *antecedent–anaphor* relation and the *cataphor–postcedent* relation, see Example 209 and Example 210, respectively.

(209) *He looked at **Magda**. The **child's** eyes, wide with amazement…*

(210) *Then he recognized **them**... Two hours ago, the **beauty** from Chicago had sat on the seat. While the black **man** in livery had gone into Kapino's for beer.*

Typically, two types of coreference relations are recognized, namely grammatical and textual. Referring expressions in grammatical coreference can be identified using grammatical rules. On the other hand, textual coreference is identified using context. The former type usually occurs with both referring expressions in the same sentence, while the latter often crosses sentence boundaries. Bridging anaphora is a relation between nominal groups where the anaphor does not directly refer to same antecedent, but an indirect relation is implied. Usually, readers can identify this relation using world knowledge and the context of discourse. For illustration, some common sense knowledge has to be used to recognize a relation between *dress* and *shoes* in Example 211. More on coreference is in Chapter 3 and on bridging anaphora in Chapter 4.

(211) *The young lady in the white **dress** was biting into a chicken leg. Yes, beside it in the grass a pair of white **shoes** had been casually tossed.*

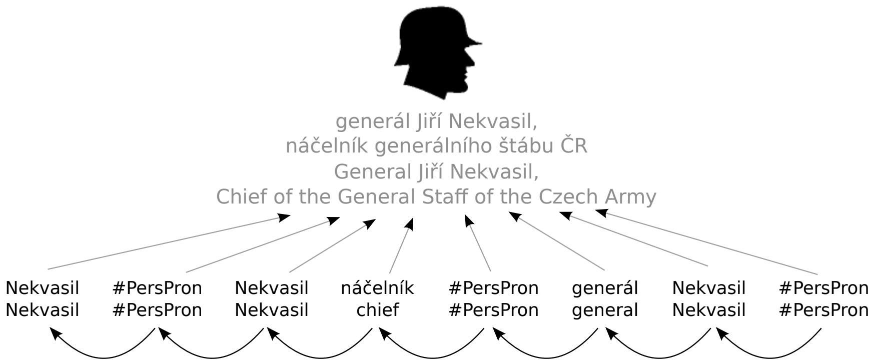


Figure 14.5: Coreference chain of *the general Jiří Nekvasil, the chief of the General Staff of the Czech army*. It consists of the eight referring expressions, i.e. the seven coreference links.

A *coreference chain* is a list of the object's referring expressions in the text. For example, the referring expressions *Magda*, *they*, *they*, *Magda*, *child's*, *they* in our 26-sentence long sample document (see Section 14.1 in this chapter) form the coreference chain for the object Magda. At the same time, each coreference chain corresponds to exactly one element in the stock of shared knowledge, see Figure 14.2.

In the PDT, coreference relations, i.e. grammatical coreference, textual coreference, and bridging anaphora, were annotated on top of tectogrammatical trees as illustrated in Figure 14.4. The referring expressions in Example 212 are highlighted in bold and their subscripts distinguish the objects they refer to. The arrow symbols in the tree are called *coreference links* and they visualize coreference relations. From this perspective, we modify the notion of referring expression in coreference relations – it is a tectogrammatical node that bears a tectogrammatical lemma referring to the real-world object. Then a coreference chain is a list of the object's referring expressions in the tectogrammatical trees of the text. An example of such a coreference chain is shown in Figure 14.5.

(212) *Olympijský vítěz v desetiboji Robert* ***Změlík***$_1$ *se v minulých dnech nastěhoval se* ***svou***$_1$ *přítelkyní Andreou Sollárovou do nového bytu na sídlišti v Praze-Řepích. Zítra* [***on***$_1$] *vyrazí do* ***francouzského***$_2$ *střediska ve Font* ***Romeu***$_2$ *k závěrečné přípravě na mistrovství světa ve Stuttgartu.* (PDT)

The Olympic decathlon medalist Robert ***Změlík***$_1$ *and* ***his***$_1$ *girlfriend Andrea Sollárová have recently moved to a new flat in a condominium in Prague-Řepy. Tomorrow,* [***he***$_1$] *will depart for the* ***French***$_2$ *resort of Font* ***Romeo***$_2$ *for a final training before the World Championship in Stuttgart.*

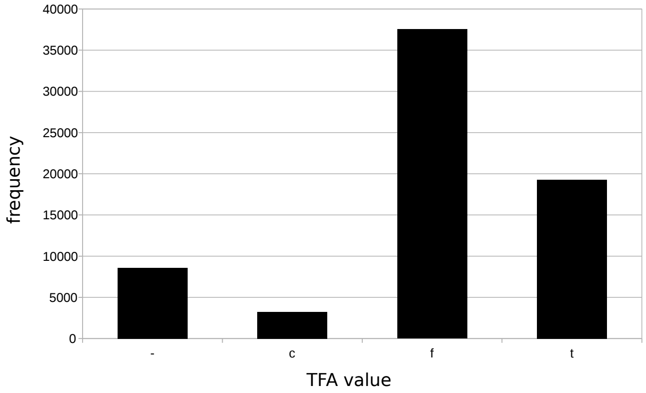


Figure 14.6: Frequency of TFA values in *train-1*

To understand the model of salience we work with, it is useful to explore the coreference chains in the PDT. The number of grammatical and textual coreference links in *train-1* and *train-all* are summarized in Table 14.1 along with the number of bridging anaphora links. Another interesting figure is the number of coreference chains.

We have adopted the definition of *coreference chain length* as the number of referring expressions present in a coreference chain. Thus studying the plot in Figure 14.7, we can see that the chains of length two are the most frequent chains in the data, whereas the frequency of longer chains is rapidly decreasing. Although the tail of the plot was cut off for the sake of readability, the longest chain encountered in the data is 89 referring expressions long (found in the document containing 114 sentences). To complete the figures, we add that the average coreference chain length is 3.26 in *train-1* and 3.25 in *train-all*.

The coreference chains are the building stones of the salience algorithm. If we want to use salience to model the dynamics of some inherent topics in the text, we need coreference chains "as long as possible." In other words, one should make an effort to identify as many coreference links as possible. Given this perspective, we also carried out some analysis on bridging anaphora links. The experimental approach is quite straightforward: since the salience algorithm does not distinguish the types of coreference relations, we can treat the bridging anaphora relations in the same way as the "regular" coreference links. However, one has to bear in mind that they do not have such "strict" characteristics, which can, to a certain degree, also affect the results of the subsequent salience modeling. For instance, if a coreference chain contains more than one bridging anaphora links, there is no guarantee that the corresponding elements in the stock of shared knowledge would remain the same throughout the whole chain without a possible semantic shift.

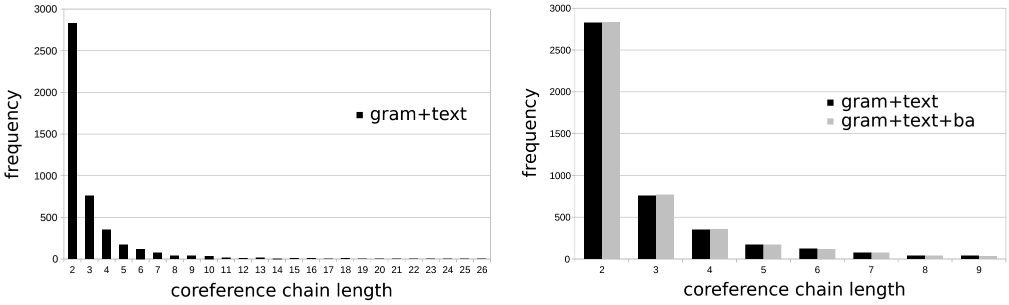


Figure 14.7: Frequency of coreference chain lengths in *train-1*

Furthermore, we examine the changes in the proportion of different coreference chain lengths when the bridging anaphora links are included. Since we expected more significant changes than those displayed in Figure 14.7, we decided not to work with the bridging anaphora links at this stage and leave them for further investigation.

Topic–focus articulation

Topic–focus articulation is one of the key notions of the Functional Generative Description framework (Sgall, Hajičová and Panevová, 1986) and stands basically for partitioning the sentence into two segments with different communicational functions. In the topic part of the sentence, the speaker mentions *what he is talking about*, while the focus part contains new information about the topic, i.e. *what he wants to say about it*. Example 213 illustrates the topic–focus segmentation, the focus part is highlighted in bold.

(213) *The young lady in the white dress* ***was biting into a chicken leg****.*

In the PDT, topic–focus articulation (TFA) was annotated on top of tectogrammatical trees as illustrated in Figure 14.4. Each relevant tectogrammatical node was assigned one of the three values: (i) *t*, a non-contrastive contextually bound node, (ii) *c*, a contrastive contextually bound node, (iii) *f*, a contextually non-bound node. The proportion of these values in *train-1* is visualized in Figure 14.6. The semantic view represented by the contextual boundness and non-boundness serves as a basis for inferring the topic–focus dichotomy and a possible segmentation of a sentence. For more on topic–focus articulation, see Chapter 5.

14.4 The Salience Algorithm

The very first salience algorithm is a deterministic knowledge-based algorithm measuring changes in the stock of shared knowledge based on the phenomena of topic–

1.	$\mathrm{dg}_r^n(x) = 0$	if r carries TFA value *f* in the n-th sentence
2.	$\mathrm{dg}_r^n(x) = -1$	if r carries TFA value *t* or *c* in the n-th sentence
3.	$\mathrm{dg}_r^n(x) = \mathrm{dg}_{r'}^{n-1}(x) - 2$	if no r is included in the n-th sentence and r' has been mentioned in the focus of the last (not necessary immediately) preceding sentence ($(n-1)$-th through 1st sentence)
4.	$\mathrm{dg}_r^n(x) = \mathrm{dg}_{r'}^{n-1}(x) - 1$	if no r is included in the n-th sentence and r' has been mentioned in the topic of the last (not necessary immediately) preceding sentence ($(n-1)$-th through 1st sentence)

Table 14.2: The salience algorithm

focus dichotomy and coreference, see Hajičová and Vrbová (1982) and Hajičová (1993). Formally, we consider the situation when an object x represented by the referring expression r has salience $\mathrm{dg}_r^n(x)$ after the n-th sentence of a document is uttered. After each sentence, the salience degree of the object x is modified, see Table 14.2.

However, the algorithm uses the notion of focus through the TFA annotation of contextual boundness *f* and likewise topic through the *c* or *t* annotation. The reason for this is rather technical: although a heuristic algorithm to assign topic and focus proposed by Sgall, Hajičová and Panevová (1986) has been formulated and tested for mapping the *c*/*t*/*f* values to the topic/focus values in Hajičová, Havelka and Veselá (2005), its results were not good enough.

Thus, the salience algorithm assigns degree of salience to the members of the coreference chains with respect to their TFA value of contextual boundness only. It concerns the rules no. 3 and 4 where mentioning of a referring expression in either focus or topic is inquired.

In Hajičová, Hladká and Kučová (2006), the algorithm was evaluated on one document only, because no other data with the necessary annotation were available at that time.

Tracing salience in the PDT

Figure 14.8 presents a salience graph for Example 214. Each coreference chain is represented by a numbered polyline and its members are marked by the corresponding color in the document.

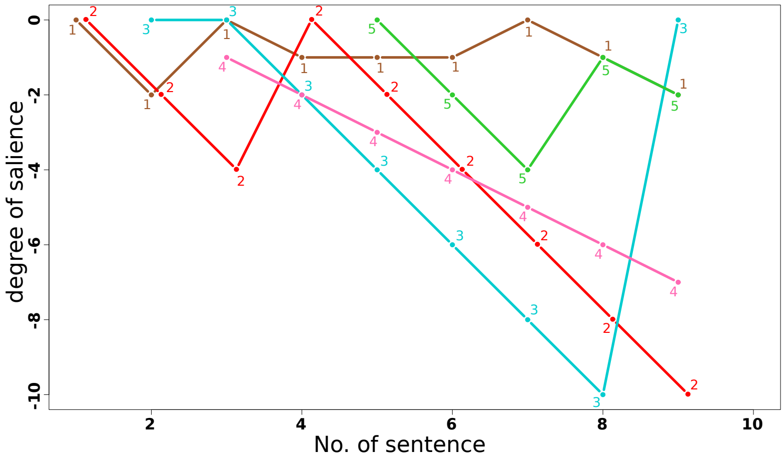


Figure 14.8: The salience graph for Example 214

(214) *Both accountant$_1$ and one million$_2$ have disappeared*
Brno$_3$
Since 11 June$_4$, when$_4$ [he$_1$] left work around 3 AM and did not come home, the police have been searching for Štefan Mišík$_1$, 27, the main accountant of casino 777 on Svobody Square in Brno$_3$. The wanted man$_1$ had over a million$_2$ crowns with him$_1$ and could be a victim of a violent crime. Štefan Mišík$_1$ resides in Pradlačka street and has short brown hair and a pea-sized birthmark on the left side of his neck$_5$, 178cm tall, medium build. [He$_1$] speaks with an accent in which the sound r is trilled. Last time [he$_1$] was wearing [on him$_1$] a bright shirt, black jeans and brown loafers. On the neck$_5$, [he$_1$] was wearing a silver chain with a Cancer zodiac sign, in a black bag he also had a new passport and cassette tapes. Witnesses can report to the nearest police office, the 158 (phone) line or the first department of Crime Service in Brno$_3$, phone 05/4116 2525. (PDT, translated)

Vertical cuts in salience graphs

Moving along the horizontal axis, i.e. reading sentence by sentence, and tracing the current trend of all the chains at once, specific vertical breaks can be identified in the salience. These breaks can signal a topic shift in the particular sentence, where several new objects emerge or are re-activated and the old ones fade away. From this point of view, the salience can be used for an automatic segmentation of a text by "cutting" it at these breaks.

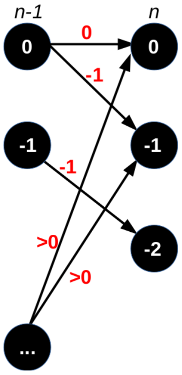


Figure 14.9: Possible values of *LeapHeight* (displayed in red by the arrows) after reading the n-th sentence: *LeapHeight* $= -1$, *LeapHeight* $= 0$, *LeapHeight* > 0

Horizontal cuts and leap height in salience graphs

Another way to approach salience would be to draw one or more horizontal lines in the graph to mark a certain degree of salience. One can assume that these degrees can express the amount of activation that an object must have to be referred to by certain grammatical means, e.g. a weak or a zero pronoun is expected to refer to an object with high activation, whereas less salient objects are re-activated by more specific expressions, e.g. a definite noun phrase. To verify these hypotheses, we introduce a new measure *salience leap height*, called *LeapHeight*.

Each time an object is mentioned in a sentence, the leap height value indicates the difference of its current salience level and its level in the previous sentence. More rigorously, the leap height value of an object x in the n-th sentence can be defined as follows:

$$\mathtt{LeapHeight}(x, n) := dg^{n}(x) - dg^{n-1}(x)$$

Note that this definition, visualized in Figure 14.9, contains not only the “depth” from which the object emerges, but it also takes into account the TFA value of the referring expression in the form of its current salience degree being either 0 or -1. This reflects the function or position of the referring expression in the current sentence. This distinction is proportionally more important with the smaller leap heights and loses its importance with their higher values, which may not necessarily be harmful.

This property also results in the leap height being zero, or even a negative number, specifically -1. If the previous referring expression of x was in the focus (i.e. having the TFA value *f*) the current referring expression is in the topic (having the TFA value *t* or *c*). This situation is actually quite common in the discourse. It corresponds

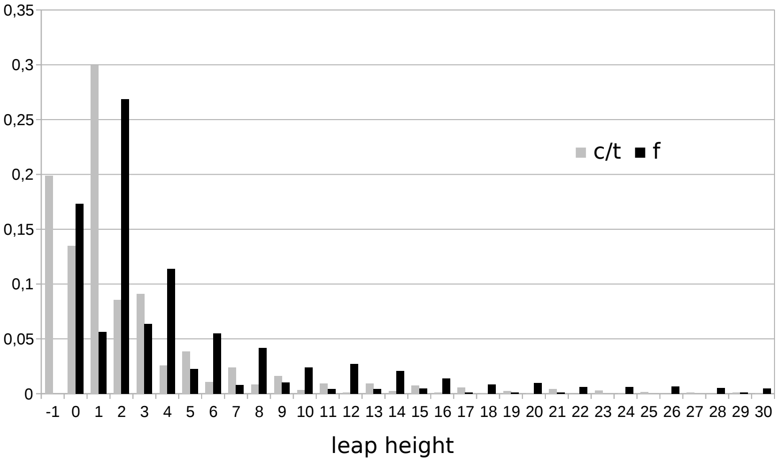


Figure 14.10: Proportion of the leap heights with respect to the referring expressions' TFA value in *train-1*. The y-axis units are the normalized ratios of leap heights for a given semantic part of speech.

to the typical situation of a newly emerged object in the $(n-1)$-th sentence that is subsequently referred to in the n-th sentence. Example 215 illustrates this situation – the object *Ministerstvo hospodářství* [*Ministry of Finance*] is referred to in each of the two sentences. The first referent is contextually non-bound (marked by *f*), thus bearing salience value 0 (indicated by the subscript). Subsequently, its occurrence in the second sentence is non-contrastive and contextually bound (the TFA value *t*), gaining the salience degree -1, which results in *LeapHeight*(*Ministerstvo hospodářství*, 2) $= -1 - 0 = -1$.

(215) *Zkušenosti* ***Ministerstva***$_{f0}$ *hospodářství ČR z loňského roku ukazují, že vzhledem k postupnému zlepšování informovanosti podnikatelů o programech podpory se podstatně zvýšil i jejich zájem o získání finančních dotací od státu. Výsledkem byl značný převis poptávky nad celkovými možnostmi, tedy prostředky, které* ***Ministerstvo***$_{t-1}$ *hospodářství dávalo k dispozici podnikatelům prostřednictvím Českomoravské záruční a rozvojové banky.* (PDT)

The experience of ***Ministry***$_{f0}$ *of Finance of the Czech Republic from last year shows that due to a gradual improvement of awareness of businessmen about support programs, their interest in public financial grants has grown substantially as well. The result was a considerable excess of demand beyond the capabilities, or resources, which the* ***Ministry***$_{t-1}$ *of Finance made available to businessmen via Czech-Moravian Guarantee and Development Bank.*

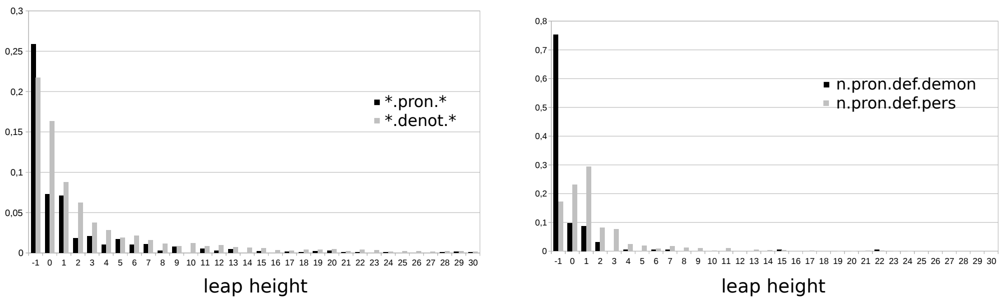


Figure 14.11: Proportion of the leap heights of indefinite pronominal semantic nouns **.pron.** and denominating semantic nouns **.denot.** (on the left) and demonstrative pronouns *n.pron.def.demon* and personal pronouns *n.pron.def.pers* (on the right) in *train-1*. The units of y-axis are normalized ratios of leap heights for the given category.

Leap Heights and TFA in salience graphs

Figure 14.10 demonstrates the proportion of the leap heights depending on the TFA value of the referring expressions. A general rule can be formulated that shorter leaps are typical for a mention in topic (c/t), while the longer ones are slightly more common for a mention in Focus (f). We should also note that the leaps to the topic are apparently more frequent for the odd leap heights, whereas the focus "destination" favors the even leap heights. This is an inherent property stemming from the inclusion of TFA in the definition of the leap height.

Pronominal vs. denominating referents

Let us return to the above mentioned hypothesis about a grammatical form of referring expressions typical for certain salience degrees. Thanks to an elaborate system of the tectogrammatical layer annotation, we can use the tectogrammatical node attribute *sempos*.[160] The pronominal expressions are marked with the *sempos* values for an indefinite pronominal semantic noun (*n.pron.indef*) and a denominating semantic noun (*n.denot*). The other values are only quantitative expressions and verbs. Then we visualize the proportion of the leap heights for each *sempos* value in Figure 14.11.[161] It is obvious that there is some disproportion in the pronominal referring expressions in comparison to the denominating ones. The quick drop of the pronominals' values beyond the leap height of 1, along with the rather steady decline of the denominators, seems to confirm the declared hypothesis. However, the dominance of the -1

[160] The sempos attribute (semantic part of speech) contains the information regarding the membership of a complex node in a semantic part of speech, for more details see Mikulová et al. (2006).

[161] Although the leap height values go as far as 172, the tail is long and its values are negligible for our research. Thus the charts are often cut off at the leap height value of 30.

value is quite surprising. Thus, Figure 14.11 focuses on comparing demonstrative and personal pronouns only (the *sempos* values *n.pron.def.demon* and *n.pron.def.pers*,[162] respectively) because these two values are by far the most frequent ones among the pronominal referring expressions.

The difference between them is apparent: while the demonstrative pronouns almost fail to refer beyond the leap height of 1 and serve mostly for the −1-leap reference, the personal pronouns, although also "specialized" on the low leaps, perform best for the leaps of 1 or 0. From this comparison, it is also evident that demonstrative pronouns are almost fully responsible for high leap height values in comparison to the leap height −1 of pronominals.

14.5 Learning Salience

The salience algorithm is a knowledge-based algorithm for modeling salience that requires an input text to be enriched with both coreference and TFA value either manually or automatically in concordance with the PDT tectogrammatical layer. Since manual annotation is a very demanding task, we evaluated performance of relevant automatic procedures to find out whether they can be employed. Mainly, we are interested in the procedure of automatic coreference resolution. For Czech, such procedure operating on the PDT tectogrammatical layer exists but its performance is relatively low, see Nedoluzhko, Mírovský and Novák (2013).

Václ (2015) conducted a pilot study to model salience by means of a supervised machine learning technique that uses morphological and analytical annotation based on the PDT framework.[163] Such a point of view can grossly oversimplify the task of modeling salience in discourse. However, the main ambition was to verify whether *LeapHeight* is an appropriate measure of changes in salience.

The study was performed with analytical trees where the number of nodes corresponds to the number of tokens in the sentence and no coreference links are resolved, see Figure 14.12. Part of speech classes of the words that have their counterparts in some coreference chain member (leaving bridging anaphora aside) are presented in Table 14.3 where the statistics are provided for both the complete PDT and its parts containing the documents of selected genres (see Chapter 2, Section 2.7.3). There is, however, one obstacle when analyzing the numbers in the table: coreference chains are defined as a sequence of coreference links between tectogrammatical nodes of tectogrammatical trees. Not all tectogrammatical nodes have their counterparts in the analytical tree, like the two rectangle nodes in the tectogrammatical tree in Figure 14.12. Such situations are denoted as *unknown*. Mostly, these tectogrammatical nodes correspond to technically added nodes to fill in a valency frame of a governing

[162] See Mikulová et al. (2006).

[163] For Czech, the automatic procedures doing such annotation, i.e. taggers and parsers, are known to have relatively high accuracy, approximately 96%, see Spoustová (2008), and 86%, see Holan and Žabokrtský (2006) and Koo et al. (2010).

Part-of-speech	All genres		Selected genres	
class	Colored	Non-colored	Colored	Non-colored
noun	9,209 (37.4%)	15,433	7,136 (38.7%)	11,319
pronoun	2,682 (70.1%)	1,146	1,880 (69.8%)	815
unknown	2,792 (23.8%)	8,944	1,879 (23.2%)	6,208
adjective	683 (7.2%)	8,801	528 (7.5%)	6,558
verb	345 (4.2%)	7,921	250 (4.1%)	5,801
adverb	178 (4.6%)	3,701	120 (4.1%)	2,785
numeral	79 (3.3%)	2,350	48 (2.9%)	1,594
other	319 (7.3%)	4,046	234 (7.3%)	2,981

Table 14.3: Amount of words of different part-of-speech classes that do (not) have their counterparts in some coreference chain member at the tectogrammatical layer of the PDT – see "colored" ("non-colored").

node. If they are involved in a coreference link, they usually present a short-range grammatical coreference link. They could not be captured in the study. On the other hand, this is not a great loss. Based on this analysis, nouns and pronouns were considered as objects to experiment with. In Figure 14.12, the nouns *letech* (*years*), *požadavky* (*request*), *členů* (*of_members*), *stávkou* (*by_strike*) and the pronoun *svých* (*its*) are considered as the objects.

Although the preliminary results of the pilot study are not fully persuasive, we can say that they look promising enough and other experiments using a different machine learning algorithm could bring improvement of our current results. In addition, experiments employing features from the tectogrammatical layer should be performed to get a rigorous comparison of knowledge-based and machine learning based approaches to modeling salience.

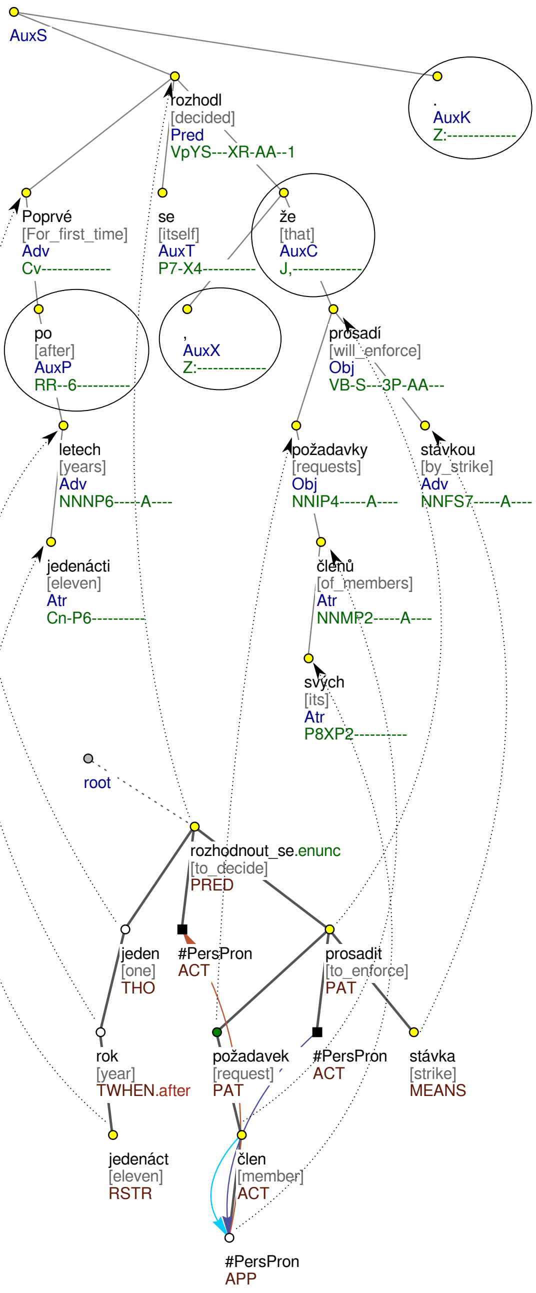


Figure 14.12: Mapping between the nodes of analytical (above) and tectogrammatical (below) trees in the PDT

15

Summary

In the analysis presented in this monograph we followed the interplays of several important coherence factors: discourse relations, coreference, bridging relations and information structure (topic–focus articulation). For this purpose, several kinds of annotation in the Prague Dependency Treebank have been linked up.

General background

The introductory part of the book presents the theoretical background of the research and the analyzed data. The important source of the research is the Functional Generative Description of the language analyzing relations between forms and functions in a structure of language layers which resulted into deep syntactic (tectogrammatical) analysis of a sentence in the Prague Dependency Treebank. Besides the Functional Generative Description, the analysis of *discourse relations* (along with their connectives, arguments and semantic classification) in the PDT was inspired by the lexical approach of the Penn Discourse Treebank (Prasad et al., 2008). In both the PDTB and the PDT, the research is based on the identification of (primary and later also secondary) discourse connectives, localization of two discourse segments they relate (discourse arguments) and assigning a semantic type to the relation. In addition, special attention was paid to other discourse phenomena in the Prague Dependency Treebank, like genres of the corpus texts or text organizing devices.

Coreference is another indicator of textual coherence addressed in the book. We distinguish between anaphoric and coreference relations, focusing on coreference and defining it as a referential identity of expressions in the text. We further distinguish grammatical and textual coreference and annotate both in the Prague Dependency Treebank, including coreference of reconstructed elided items. Within textual coreference, coreference relation of specific and generic nominal groups are discriminated.

Bridging relations are coherence-relevant relations between nominal groups which go beyond the notion of coreference. We describe them as an inference about non-coreferential expressions introduced in a text which are related in some particular way that is not explicitly stated, but this relation contributes essentially to the text coherence. In the PDT, we distinguish six types of bridging relations: part–whole, set–subset, object–function, contrast, explicit anaphora without coreference and an underspecified group REST.

Another aspect of text coherence is expressed by the organization of the *topic–focus articulation*. In the PDT, the topic–focus articulation is annotated according to the theory of the Functional Generative Description. The fundamental features which the Functional Generative Description works with are *contextual boundness* and *communicative dynamism*. These two phenomena also serve as the basis for delimitation of topic and focus. At the same time, topic and focus may be well detectable according to two operational criteria – the question test and test with negation. In the corpus, topic–focus articulation is annotated on the tectogrammatical (deep syntactic) layer of language from two perspectives – as contextual boundness and communicative dynamism in dependency trees. The division of sentences into topic and focus is not explicitly marked but it is clearly deducible from the annotation of contextual boundness and communicative dynamism.

Data

This theoretical basis was applied to the annotation of different kinds of text relations in the Prague Dependency Treebank, a corpus with multi-layer annotations of Czech journalistic texts. The layers of annotation of the PDT are described, to prepare the reader for the subsequent case studies, which were carried out on the data of the PDT. We show the importance of measuring the inter-annotator agreement during all stages of any corpus annotation and demonstrate on the example of, primarily, the Prague Dependency Treebank that the inter-annotator agreement gets generally lower as we go deeper in the layers of language description, from morphology to the sentence structure and to the text structure, i.e. from relatively simple surface phenomena to more complex, vague and dubious phenomena. As the PDT presents an extraordinarily rich source of annotated data, short instructions for searching in the PDT are presented, introducing a powerful, yet easy-to-use and intuitive PML-Tree Query system. On a series of examples, it is demonstrated not only how to search for individual results of the queries in the data, but also how to utilize a system of *output filters* to produce complex summaries of all the occurrences of the query results in the data.

Case studies

Our analysis of relations in a text proceeded from single specific research areas such as discourse relations, coreference etc. to more complex questions concerning the interplay of these perspectives. From structural layers on one side, such as syntax and tectogrammatics, and from single words (discourse connectives, demonstrative pronouns) on the other side we arrived at the study of the inter-relations of their functions, concerning e.g. the stock of shared knowledge between the speaker and the recipient and the salience of different elements in a text. First, we concerned the *relation between discourse structure and syntax*. Our research on discourse structure is based on the analysis of syntactic and semantic relations which were examined on the tectogrammatical layer of the corpus. After completing the annotation of tectogrammatics, our research documented that the annotated data contain information not only about the sentences themselves but also about relations between them and, in fact, about a complex structure of the whole text. Thus, by analyzing discourse structure, we could take advantage of the richness of the syntactic and semantic annotation.

In particular, discourse annotation carried out on dependency trees in the PDT exploited the following features of syntactic analysis: (i) in 80% of intra-sentential discourse-relevant kinds of syntactic relations between clauses, syntactico-semantic labels could be transferred to the discourse annotation; (ii) in 99% of cases, the syntactic structure captured by tree-defined scope of discourse argument could be used (this feature was helpful for both intra-sentential and inter-sentential relations); (iii) connectives of intra-sentential discourse relations could be found in places predefined by syntactic analysis; (iv) coordination resolution made the discourse annotation more comprehensible by defining the structure of sentences especially in syntactically complicated cases; (v) ellipsis resolution enables the annotation of structures with verb elided in the surface form of the sentence. Thus, after manual annotation of places where syntactic analysis differs from discourse analysis, we were able to extract all of the mentioned information automatically. Analysis of places where syntactic annotation does not correspond to discourse annotation confirmed the assumption that these two types of analysis differ especially with regard to semantics – while syntactic analysis considers form and meaning as equally important features of structures, discourse analysis tracks meaning systematically disregarding the form. Furthermore, discourse analysis goes beyond sentence boundary and syntactic analysis thus cannot offer any clues for it in these cases.

However, discourse analysis carried out on syntactically annotated data enables complex comparison of discourse and syntax analysis of text. Thus, it could be shown that intra-sentential relations are more numerous than relations between separate sentences (approx. 12,600 intra-sentential versus 5,500 inter-sentential). Furthermore, among relations realized within a single sentence, coordinate structures are predominant (70% of all intra-sentential realizations), while subordinate constructions seem to realize discourse relations rather marginally. Moreover, it was possible to establish scale of individual semantic discourse types according to their realization both within a single sentence versus between sentences and in coordinate versus subordinate structures.

One of the first tasks of discourse-oriented research was to delimit the category of *discourse connectives*, similarly as in the Penn Discourse Treebank. There were two steps in setting up the category, resulting in the division of expressions with the connecting function into two subcategories. It is the so-called *primary discourse connectives* (coordinate and some subordinate conjunctions, some particles and adverbs; the group roughly corresponds to discourse connectives annotated in the PDTB) and *secondary discourse connectives* (longer and less fixed connective phrases, similar to the category of alternative lexicalizations annotated in the PDTB).

In comparison to English, Czech is a language with rich inflection and free word order; these typological differences naturally also influence the criteria for discourse connective category delimitation and the underlying language specific features have to be addressed. For primary connectives, the part-of-speech appurtenance was discussed followed by the issue of their (in)flectibility, historical development of the group and their possible placement in Czech clauses and sentences. Second, we present basic distributions of primary connectives annotated in the Prague Dependency Treebank 3.0, pointing out morphosyntactic and semantic properties of those most frequent.

Secondary connectives as well as primary connectives have connecting function in the text, i.e. they reflect semantic relations between discourse units (arguments). However, the form of secondary connectives is mostly not fixed. In comparison with primary connectives, they represent a very variable class of expressions. For example, secondary connectives are mostly modifiable (e.g. *the main/important reason is*). According to PDT data, Czech secondary connectives are realized mostly by verbal phrases, prepositional phrases or (semi-)clauses. The heterogeneity of secondary connectives is seen also in their lexical characteristics – they form a scale between lexically free and lexically frozen expressions. Semantically, they contain either explicit or implicit anaphoric reference to the previous discourse argument (cf. *the example of this is* vs. *the example is*). The analysis of PDT data has demonstrated that as for frequency of occurrences, secondary connectives constitute a minority of discourse connectives. Also in this sense, the term secondary is suitable for them.

Since the coherence is a phenomenon established by different means on many levels, often redundantly, we tried to explicate it from the opposite perspective searching for *weak coherence* in a text. This search was carried out in two experiments. In the first case, we focused on pairs of sentences between which none of the annotated relations can be observed, be it expressed with a discourse connective (primary or secondary), coreference expressions or with means of the topic–focus articulation. The result was twofold: (i) further types of clearly delimited relations were found which have been put aside by the annotation so far (*attribution*, the relation between *text organizing devices* like authors´ name or title of the text and the basic text; the relation between *question and answer*); (ii) the reader's expectation of meeting certain rules was emphasized as an important coherence factor (e.g. expectation of thematic unity, pro-coherent interpretation of ellipses). A text genre and its way of segmenting typical information into an assumed structure turned out to be a substantial feature cooperating in creating coherence. On the other hand, some coherence factors proved to be vague, although undoubtedly present; their unambiguous and clearly defined marking in an annotation would probably be problematic. The occurrence of such cases caused the seeming coherence disruptions in our data: in fact, these instances are not disruptive, but the ways of connecting the segments have not been clearly defined so far (cf. e.g. connection based on the thematic unity).

In the second experiment concerning incoherence, we followed the interplay of the topic–focus articulation, coreference and bridging relations. We focused on non-contrastive contextually bound expressions which have no antecedent in the previous context. The question arose about what their contextual boundness is based on if their antecedents do not occur in the text. The analysis showed that disregarding a number of more or less technical reasons indicated by the tectogrammatical structure of PDT sentences and annotation errors, the reasons are the following: (i) some contextually bound nominal groups are semantically or pragmatically related to previous textual or extralinguistic context but not specified as bridging relations within the PDT; (ii) the nominal groups under discussion express secondary circumstances (temporal, local, etc.), and (iii) some nominal groups have low referential potential.

As in the case of the study of weak coherence between sentences, a practical restriction of the annotation played its role in this case, too. It turns out that a sufficiently general and leisurely approach to the text interpretation allows to interconnect many elements in the text almost without any restriction, tolerating the significant variation and vagueness of expressing coherence. The question is how all of these ways can be captured reliably and objectively in an annotation, so that an acceptable inter-annotator agreement is achieved. We seem to have come across a border in the analysis, namely a border of annotation possibilities. The ways of expressing coherence relations may be redundant and variable due to the fact that the perception of a text is variable and subjective. A question arises then how such phenomena should be captured.

Finally, our complex analysis of the interplay of the discourse related topics led to us carrying out an analysis of the stock of shared knowledge between the speaker and the hearer, namely the activation of its elements in a text. We tried to "read out" as much information as possible from the sentence underlying structure represented in the form of a dependency tree, the topic–focus articulation of the sentence, and coreference relations. In other words, we worked with the notion of the *degree of salience* of the items in the stock of shared knowledge together with the representation of the dynamic development of the discourse by means of changing these degrees.

We applied machine learning techniques to model the rule-based salience algorithm formulated earlier, along with a visualization of its results. This was achieved using the PDT. A notion of salience leap height was introduced and used to explore the possibility to predict the salience degree automatically. The results of these pilot experiments were quite positive, although they cannot be used as a feature in natural language systems (e.g. machine translation) yet. Other experiments must be performed in order for this to happen.

The analysis performed on large-scale corpus data proved that coherence is ensured at various levels, from individual words, such as text connectors or demonstratives over sentence structure to the structure of the whole text which is e.g. linked by coreference chains but also resembles certain genre construction. Single aspects of the text coherence, like discourse relations, coreference, bridging relations and topic–focus articulation cooperate building the general net of connections. The result is a complex structure which aims to keep the reader oriented in the text and to relay to him the general sense of the text as a whole.

List of Abbreviations

a-layer, *analytical layer*
a-node, *analytical node*
Acc, *accusative*
ACMP, *Accompaniment (functor)*
ACT, *Actor (functor)*
ADDR, *Addressee (functor)*
Adv, *Adverbial (analytical function)*
ADVS, *Adversative relation (functor)*
afun, *analytical function (a-layer attribute)*
AIM, *Aim (functor)*
AltLex, *alternative lexicalization of a discourse connective*
ANAPH, *non-coreferential anaphoric relation (bridging relation)*
ANNIS, *Annotation of Information Structure*
APP, *Appurtenance (functor)*
APPS, *Apposition (functor)*
Arg1, *argument 1 of a discourse relation*
Arg2, *argument 2 of a discourse relation*
Atr, *attribute (analytical function)*
AUTH, *Author (functor)*

BioDRB, *Biomedical Discourse Relation Bank*

c, *contrastive contextually bound item*
CAUS, *Cause (functor)*
CD, *communicative dynamism*
CNCS, *Concession (functor)*
COMPL, *Complement (functor)*
COND, *Condition (functor)*
CONFR, *Confrontation (functor)*
CONJ, *Conjunction (functor)*
CONTRD, *Contradiction (functor)*
#Cor, *controllee in control constructions (tectogrammatical lemma)*
coref_special, *special type of coreference*
CSQ, *Consequence (functor)*

DannPASS, *Danish Phonetically Annotated Spontaneous Speech*
Db, *adverb (morphological characteristics)*
DC, *discourse connective*
DIR, *Direction (functor)*
DISJ, *Disjunction (functor)*

#EmpNoun, *non-expressed noun (tectogrammatical lemma)*
EntRel, *entity-based relation*
exoph, *exophora*

f, *contextually non-bound item*
FGD, *Functional Generative Description*
FSP, *functional sentence perspective*
FUNCT_P, *function–entity (bridging relation)*

GEN, *generic (coreference relation)*
#Gen, *general participant (tectogrammatical lemma)*
GRAD, *Gradation (functor)*
gram/sempos, *grammateme, semantic part of speech*

JCon, *coordinate conjunction (morphological characteristics)*
JJX, *adjectival phrase*
JSub, *subordinate conjunction (morphological characteristics)*

LOC, *Locative (functor)*

m-layer, *morphological layer*
MANN, *Manner (functor)*
MAT, *Material, partitive (functor)*
MorfFlex, *Czech morphological dictionary*

n.denot, *denominating semantic noun*
n.pron.def.demon, *definite pronominal semantic noun – demonstrative pronoun*
n.pron.def.pers, *definite pronominal semantic noun – personal pronoun*
n.pron.indef, *indefinite pronominal semantic noun*
NLP, *natural language processing*
Nom, *nominative*
NoRel, *no relation*
NX, *noun phrase (morphological characteristics)*

Obj, *Object (analytical function)*
#Oblfm, *obligatory adjunct (tectogrammatical lemma)*
opp, *opposition (type of discourse relation)*

P_FUNCT, *entity–function (bridging relation)*
PAR, *Parenthesis (functor)*
PART_WHOLE, *part–whole (bridging relation)*
PAT, *Patiens (functor)*
PCEDT, *Prague Czech-English Dependency Treebank*
PDiT, *Prague Discourse Treebank*
PDT, *Prague Dependency Treebank*
PDTB, *Penn Discourse Treebank*
PE, *relative pronoun což (morphological characteristics)*
PEDT, *Prague English Dependency Treebank*
#PersPron, *personal pronoun (tectogrammatical lemma)*
PML, *Prague Markup Language*
PML-TQ, *PML-Tree Query*
PoS, *part of speech*
PPX, *prepositional phrase*
PREC, *reference to preceding context (functor)*
Pred, *predicate (analytical function)*
PRED, *Predicate (functor)*

#Rcp, *participant left out as a result of reciprocation (tectogrammatical lemma)*
REAS, *Reason (functor)*
reason, *reason–result (type of discourse relation)*
RHEM, *Rhematizer (functor)*
RST, *Rhetorical Structure Theory*
RSTR, *Restricting or specifying modification (functor)*

Sb, *subject (analytical function)*
segm, *segment*
SET_SUB, *set–subset (bridging relation)*
SPEC, *specific (coreference relation)*
SUB_SET, *subset–set (bridging relation)*
SUBS, *Substitution (functor)*

t, *non-contrastive contextually bound item*
t_lemma, *tectogrammatical lemma*
t-layer, *tectogrammatical layer*
t-node, *tectogrammatical node*
TFA, *topic–focus articulation*
tfa, *topic–focus articulation (t-layer attribute)*
TrEd, *Tree Editor*
TT, *particle (morphological characteristics)*

#Unsp, *non-specific participant (tectogrammatical lemma)*

VX, *verb phrase*
w-layer, *word layer*
WHOLE_PART, *whole–part (bridging relation)*

Z, *punctuation (morphological characteristics)*

Bibliography

Adamec, Přemysl. 1980. K vyjadřování referenční určenosti v češtině a ruštině [Expression of Definiteness in Czech and English]. *Slovo a slovesnost*, 41(4), 257–264.

Arutunova, Natalia D. 1976. *Predloženije i jego smysl [Sentence and Its Meaning]*. Moscow: Nauka.

Asher, Nicholas. 1993. *Reference to Abstract Objects in Discourse*. Norwell: Kluwer Academic Publishers.

Asher, Nicholas and Alex Lascarides. 1998. Bridging. *Journal of Semantics*, 15(1), 83–113.

Asher, Nicholas and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge, UK: Cambridge University Press.

Bärenfänger, Maja, Daniela Goecke, Mirco Hilbert, Harald Lüngen and Maik Stührenberg. 2008. Anaphora as an Indicator of Elaboration: A Corpus Study. *Journal for Language Technologies and Computational Linguistics*, 15(2), 49–72.

Barzilay, Regina and Mirella Lapata. 2008. Modeling Local Coherence: An Entity-Based Approach. *Computational Linguistics*, 34(1), 1–34.

Bauer, Jaroslav. 1960. *Vývoj českého souvětí [Development of a Compound Sentence in Czech]*. Prague: Nakladatelství Československé akademie věd.

Bauer, Jaroslav. 1972. Spojky a příslovce [Conjunctions and Adverbs]. In *Syntactica slavica: Vybrané práce ze slovanské skladby [Selected Papers on Slavic Syntax]*. Brno: Universita J. E. Purkyně, 351–360.

Baumann, Stefan, Caren Brinckmann, Silvia Hansen-Schirra, Geert-Jan M. Kruijff, Ivana Kruijff-Korbayová, Stella Neumann, Erich Steiner, Elke Teich and Hans Uszkoreit. 2004. The MULI Project: Annotation and Analysis of Information Structure in German and English. In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, Raquel Silva, Carla Pereira, Filipa Carvalho, Milene Lopes, Mónica Catarino and Sérgio Barros (eds.), *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon: European Language Resources Association, 1489–1492.

de Beaugrande, Robert-Alain and Wolfgang Dressler. 1981. *Introduction to Text Linguistics*. London: Longman.

Bedřichová, Zuzanna. 2008. Částice implikující presupozici jako podstatná složka větného významu [Particles Implicating Presupposition as Significant Part of Sentence Meaning]. *Čeština doma a ve světě*, 16(3–4), 119–126.

Bejček, Eduard, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Magda Ševčíková, Jan Štěpánek and Šárka Zikánová. 2013. *Prague Dependency Treebank 3.0*. Data/software. Prague: Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics. [cit. 2015_07_22].

Bejček, Eduard, Jarmila Panevová, Jan Popelka, Lenka Smejkalová, Pavel Straňák, Magda Ševčíková, Jan Štěpánek, Josef Toman, Zdeněk Žabokrtský and Jan Hajič. 2011. *Prague Dependency Treebank 2.5*. Data/software. Prague: Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics. [cit. 2015_07_22].

Bejček, Eduard and Pavel Straňák. 2010. Annotation of Multiword Expressions in the Prague Dependency Treebank. *Language Resources and Evaluation*, 44(1–2), 7–21.

Bejček, Eduard, Pavel Straňák and Pavel Pecina. 2013. Syntactic Identification of Occurrences of Multiword Expressions in Text Using a Lexicon with Dependency Structures. In *The 9th Workshop on Multiword Expressions (MWE 2013)*. Atlanta: Association for Computational Linguistics, 106–115.

Bémová, Alevtina, Jan Hajič, Barbora Vidová Hladká and Jarmila Panevová. 1999. Morphological and Syntactic Tagging of the Prague Dependency Treebank. In *Journées ATALA – Corpus annotés pour la syntaxe; ATALA Workshop – Treebanks*. Paris: Université Paris, 21–29.

Biber, Douglas and Susan Conrad. 1999. Lexical Bundles in Conversation and Academic Prose. In Hilde Hasselgård and Signe Oksefjell (eds.), *Out of Corpora: Studies in Honour of Stig Johansson*. Atlanta: Rodopi Press, 181–190.

Bohnet, Bernd, Alicia Burga and Leo Wanner. 2013. Towards the Annotation of Penn TreeBank with Information Structure. In Ruslan Mitkov and Jong C. Park (eds.), *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. Nagoya: Asian Federation of Natural Language Processing, 1250–1256.

Brants, Thorsten. 2000. Inter-Annotator Agreement for a German Newspaper Corpus. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*. Athens: European Language Resources Association.

Calhoun, Sasha, Malvina Nissim, Mark Steedman and Jason Brenier. 2005. A Framework for Annotating Information Structure in Discourse. In Adam Meyers (ed.), *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*. Ann Arbor: Association for Computational Linguistics, 45–52.

Carlson, Greg and Francis Jeffry Pelletier. 1995. *The Generic Book*. Chicago: University of Chicago Press.

Carlson, Lynn and Daniel Marcu. 2001. *Discourse Tagging Reference Manual*. Technical report ISI-TR-545. Los Angeles: Information Sciences Institute, University of Southern California. [cit. 2015_07_22]. Available from <http://www.isi.edu/~marcu/discourse/tagging-ref-manual.pdf>.

Carlson, Lynn, Mary Ellen Okurowski and Daniel Marcu. 2002. *RST Discourse Treebank*. Data/software. Philadelphia: Linguistic Data Consortium. [cit. 2015_07_22].

Carnap, Rudolf. 1947. *Meaning and Necessity*. Chicago: University of Chicago Press.

Čermák, František. 2007. *Frazeologie a idiomatika česká a obecná [Czech and General Phraseology]*. Prague: Karolinum.

Chafe, Wallace L. 1976. Givenness, Contrastiveness, Definiteness, Subjects, Topics, and Point of View. In Charles N. Li (ed.), *Subject and Topic*. Cambridge, MA, USA: Academic Press, 25–55.

Charolles, Michel. 1999. Associative Anaphora and Its Interpretation. *Journal of Pragmatics*, 31(3), 311–326.

Cheng, Hua, Massimo Poesio, Renate Henschel and Chris Mellish. 2001. Corpus-based NP Modifier Generation. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*. Stroudsburg: Association for Computational Linguistics.

Chernejko, Ludmila O. 1997. *Abstraktnoje imja. Lingvo-filosofskij analiz abstraktnogo imeni [Abstract Nouns. Linguistic and Philosophical Analysis of Abstract Nouns].* Moscow: Moscow State University.

Chomsky, Noam. 1969. *Deep Structure, Surface Structure, and Semantic Interpretation*. Bloomington: Indiana University Linguistics Club.

Clark, Herbert H. 1975. Bridging. In David Waltz (ed.), *Theoretical Issues in Natural Language Processing*. New York: Association for Computing Machinery, 169–174.

Cohen, Jacob. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1), 37–46.

Daneš, František. 1974. *Papers on Functional Sentence Perspective*. Prague: Academia.

Daneš, František. 1979. O identifikaci známé (kontextově zapojené) informace v textu [Identification of Known (Contextually Bound) Information in Text]. *Slovo a slovesnost*, 40(1), 257–270.

Daneš, František. 2009. Takzvané „vztažné věty nepřívlastkové“: současné názory na jejich status [So-Called False Relative Clauses: Present-day Opinions on Their Status]. *Naše řeč*, 92(4), 169–183.

Daneš, František, Helena Běličová, Mirek Čejka, Emil Dvořák, Miroslav Grepl, Karel Hausenblas, Zdeněk Hlavsa, Jana Hoffmannová, Josef Hrbáček, Jan Chloupek, Petr Karlík, Eva Macháčková, Olga Müllerová, Bohumil Palek, Jiří Nekvapil, Jiří Novotný, Petr Piťha, Hana Prouzová, Milena Rulfová, Blažena Rulíková, Otakar Šoltys, Ludmila Uhlířová and Stanislav Žaža. 1987. *Mluvnice češtiny. 3. Skladba [Grammar of Czech. 3. Syntax].* Prague: Academia.

Dipper, Stefanie, Michael Götze and Stavros Skopeteas. 2007. *Information Structure in Cross-Linguistic Corpora: Annotation Guidelines for Phonology, Morphology, Syntax, Semantics and Information Structure*. Potsdam: Universitätsverlag Potsdam.

Firbas, Jan. 1964. On Defining the Theme in Functional Sentence Perspective. *Travaux Linguistique de Prague*, 1(1), 267–280.

Firbas, Jan. 1971. On the Concept of Communicative Dynamism in the Theory of Functional Sentence Perspective. *Brno Studies in English*, 7(19), 12–47.

Firbas, Jan. 1992. *Functional Sentence Perspective in Written and Spoken Communication*. Cambridge, UK: Cambridge University Press.

Forbes-Riley, Katherine, Bonnie Webber and Aravind Joshi. 2006. Computing Discourse Semantics: The Predicate-Argument Semantics of Discourse Connectives in D-LTAG. *Journal of Semantics*, 23(1), 55–106.

Frege, Gottlob. 1892. Über Sinn und Bedeutung. *Zeitschift für Philosophie und philologische Kritik*, 100, 25–50.

Fried, Mirjam. 2011. Vztažné věty s nesklonným *co* [Relative Clauses with the Indeclinable Pronoun *co*]. In František Štícha (ed.), *Kapitoly z české gramatiky [Chapters from Czech Grammar]*. Prague: Academia, 1126–1143.

von der Gabelentz, Georg. 1868. Ideen zu einer vergleichenden Syntax – Wort- und Satzstellung. *Zeitschrift für Völkerpsychologie und Sprachwissenschaft*, 6(1), 376–384.

Gardent, Claire, Helene Mahuelian and Eric Kow. 2003. Which Bridges for Bridging Definite Descriptions? In Anne Abeille, Silvia Hansen-Schirra and Hans Uszkoreit (eds.), *Proceedings of the EACL 2003 Workshop on Linguistically Interpreted Corpora*. Budapest: European Language Resources Association, 69–76.

Gebauer, Jan. 1970. *Slovník staročeský [Old-Czech Dictionary]*. Prague: Academia.

Grepl, Miroslav and Petr Karlík. 1998. *Skladba češtiny [Syntax of Czech]*. Olomouc: Votobia.

Grosz, Barbara J. and Candace L. Sidner. 1986. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, 12(3), 175–204.

Grosz, Barbara J., Scott Weinstein and Aravind Joshi. 1995. Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, 21(2), 203–225.

Gundel, Jeanette K., Nancy Hedberg and Ron Zacharski. 1993. Cognitive Status and the Form of Referring Expressions in Discourse. *Language*, 69(2), 274–307.

Hajič, Jan. 2004. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Prague: Karolinum.

Hajič, Jan. 2006. Complex Corpus Annotation: The Prague Dependency Treebank. In Mária Šimková (ed.), *Insight into the Slovak and Czech Corpus Linguistics*. Bratislava: Veda, 54–73.

Hajič, Jan, Barbora Vidová Hladká, Jarmila Panevová, Eva Hajičová, Petr Sgall and Petr Pajas. 2001. *Prague Dependency Treebank 1.0 (Final Production Label)*. Data/software. Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Philadelphia: Linguistic Data Consortium. [cit. 2015_07_22].

Hajič, Jan and Jaroslava Hlaváčová. 2013. *MorfFlex CZ*. Data/software. Prague: Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, LINDAT/CLARIN digital library. [cit. 2015_08_30]. Available from <http://hdl.handle.net/11858/00-097C-0000-0015-A780-9>.

Hajič, Jan, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, Magda Ševčíková-Razímová and Zdeňka Urešová. 2006. *Prague Dependency Treebank 2.0*. Data/software. Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Philadelphia: Linguistic Data Consortium. [cit. 2015_07_22].

Hajič, Jan, Jarmila Panevová, Zdeňka Urešová, Alevtina Bémová, Veronika Kolářová and Petr Pajas. 2003. PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In Joakim Nivre and Erhard Hinrichs (eds.), *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*. Vaxjo: Vaxjo University Press, 57–68.

Hajičová, Eva. 1972. Some Remarks on Presuppositions. *The Prague Bulletin of Mathematical Linguistics*, 17, 11–23.

Hajičová, Eva. 1973. Negation and Topic vs. Comment. *Philologica Pragensia*, 16(1), 81–93.

Hajičová, Eva. 1975. *Negace a presupozice ve významové stavbě věty [Negation and Presupposition in Semantic Structure of Sentence]*. Prague: Academia.

Hajičová, Eva. 2012. Topic–Focus Revisited (Through the Eyes of the Prague Dependency Treebank). In Jurij D. Apresjan (ed.), *Smysly, teksty i drugie zachvatyvajuščie sjužety. Sbornik statej v čest 80-letija Igorja Aleksandroviča Melčuka*. Moscow: Jazyky slavjanskoj kultury, 218–232.

Hajičová, Eva, Karel Oliva and Petr Sgall. 1987. Odkazování v gramatice a v textu [Coreference in the Grammar and in the Text]. *Slovo a slovesnost*, 48(3), 199–212.

Hajičová, Eva, Petr Pajas and Kateřina Veselá. 2002. Corpus Annotation on the Tectogrammatical Layer: Summarizing the First Stages of Evaluations. *The Prague Bulletin of Mathematical Linguistics*, 77, 5–18.

Hajičová, Eva, Jarmila Panevová and Petr Sgall. 1985. Coreference in the Grammar and in the Text. Part I. *The Prague Bulletin of Mathematical Linguistics*, 44, 2–22.

Hajičová, Eva, Jarmila Panevová and Petr Sgall. 1986. Coreference in the Grammar and in the Text. Part II. *The Prague Bulletin of Mathematical Linguistics*, 46, 1–11.

Hajičová, Eva, Jarmila Panevová and Petr Sgall. 1987. Coreference in the Grammar and in the Text. Part III. *The Prague Bulletin of Mathematical Linguistics*, 48, 3–12.

Hajičová, Eva. 1993. *Issues of Sentence Structure and Discourse Patterns*. Prague: Charles University Press.

Hajičová, Eva. 2003. Contextual Boundness and Discourse Patterns. In Jiří Mírovský, Anna Kotěšovcová and Eva Hajičová (eds.), *Proceedings of XVII International Congress of Linguists, CD-ROM*. Prague: Matfyzpress.

Hajičová, Eva, Jiří Havelka and Kateřina Veselá. 2005. Corpus Evidence of Contextual Boundness and Focus. In Pernilla Danielsson (ed.), *Proceedings of the Corpus Linguistics Conference Series*. Birmingham: University of Birmingham, 1–9.

Hajičová, Eva, Barbora Hladká and Lucie Kučová. 2006. An Annotated Corpus as a Test Bed for Discourse Structure Analysis. In Candy Sidner, John Harpur, Anton Benz and Peter Kuhnlein (eds.), *Proceedings of the Workshop on Constraints in Discourse Structure Analysis*. Maynooth: National University of Ireland, 82–89.

Hajičová, Eva, Barbara Partee and Petr Sgall. 1998. *Topic–Focus Articulation, Tripartite Structures, and Semantic Content*. Dordrecht: Kluwer Academic Publishers.

Hajičová, Eva and Jarka Vrbová. 1982. On the Role of the Hierarchy of Activation in the Process of Natural Language Understanding. In Ján Horecký (ed.), *Proceedings of the 9th Conference on Computational Linguistics*. Prague: Academia, 107–113.

Hajičová, Eva. 1995. Postavení rematizátorů v aktuálním členění věty [Position of Rhematizers in the Topic–Focus Articulation]. *Slovo a slovesnost*, 56(4), 241–251.

Halliday, Michael Alexander Kirkwood. 1967. Notes on Transitivity and Theme in English: Part 1. *Journal of Linguistics*, 3(1), 37–81.

Halliday, Michael Alexander Kirkwood and Ruqaiya Hasan. 1976. *Cohesion in English*. London: Longman.

Hana, Jiří, Daniel Zeman, Jan Hajič, Hana Hanová, Barbora Hladká and Emil Jeřábek. 2005. *Manual for Morphological Annotation, Revision for the Prague Dependency Treebank 2.0*. Technical report TR-2005-27. Prague: Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics. [cit. 2015_08_30]. Available from <http://ufal.mff.cuni.cz/techrep/tr27.pdf>.

Hana, Jirka and Jan Štěpánek. 2012. Prague Markup Language Framework. In *Proceedings of the Sixth Linguistic Annotation Workshop*. Stroudsburg: Association for Computational Linguistics, 12–21.

Haviland, Susan E. and Herbert H. Clark. 1974. What's New? Acquiring New Information as a Process in Comprehension. *Journal of Verbal Learning and Verbal Behavior*, 13(5), 512–521.

Havránek, Bohuslav, Jaromír Bělič, Miloš Helcl, Alois Jedlička, Vlasta Červená, Josef Filipec, Františka Havlová, Miloslav Churavý, Ladislav Janský, Karla Kozlová, Libuše Kroupová, Jaroslav Machač, Hana Marešová, Vladimír Mejstřík, Emanuel Michálek, Blanka Papírníková, Eva Pokorná, Běla Poštolková, Miroslav Roudný, Zdeňka Sochová, Naďa Svozilová, Eva Vodrážková, Jaroslav Zima, František Daneš, Karel Hausenblas, Zdeňka Hlaváčková, Alena Hovorková, Helena Kratochvílová, Jitka Štindlová, František Váhala and Josef Vachek. 1989. *Slovník spisovného jazyka českého [Dictionary of Standard Czech]*. Prague: Academia.

Hendrickx, Iris, Orphée De Clercq and Veronique Hoste. 2011. Analysis and Reference Resolution of Bridge Anaphora across Different Text Genres. In Iris Hendrickx, Sobha Lalitha Devi, António Branco and Ruslan Mitkov (eds.), *Anaphora Processing and Applications*. Berlin/Heidelberg: Springer, 1–11.

Hlavsa, Zdeněk. 1972. K protikladu určenosti v češtině [Opposites of Definiteness in Czech]. *Slovo a slovesnost*, 33(3), 199–203.

Hlavsa, Zdeněk. 1975. *Denotace objektu a její prostředky v současné češtině [Denotation of an Object and Its Realization in Modern Czech]*. Prague: Academia.

Hobbs, Jerry R. 1979. Coherence and Coreference. *Cognitive Science*, 3(1), 67–90.

Hoffmannová, Jana. 1993. Koherence, koheze, konexe...? [Coherence, Cohesion, Connection...?]. *Slovo a slovesnost*, 54(1), 58–64.

Holan, Tomáš and Zdeněk Žabokrtský. 2006. Combining Czech Dependency Parsers. In Petr Sojka, Ivan Kopeček and Karel Pala (eds.), *Text, Speech and Dialogue*. Berlin/Heidelberg: Springer, 95–102.

Hou, Yufang, Katja Markert and Michael Strube. 2013. Global Inference for Bridging Anaphora Resolution. In Lucy Vanderwende, Hal Daumé III and Katrin Kirchhoff (eds.), *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta: Association of Computational Linguistics, 907–917.

Howarth, Peter. 1998. The Phraseology of Learners' Academic Writing. In Anthony Paul Cowie (ed.), *Phraseology: Theory, Analysis, and Applications*. Oxford: Oxford University Press, 161–186.

Howarth, Peter. 2000. Describing Diachronic Change in English Phraseology. In Gloria Corpas Pastor (ed.), *Las Lenguas de Europa: Estudios de fraseología, fraseografía y tradducción Interlingua*. Granada: Comares, 213–230.

Hrbáček, Josef. 1994. *Nárys textové syntaxe spisovné češtiny [Delineation of Textual Syntax of Standard Czech]*. Prague: Trizonia.

Jínová, Pavlína, Jiří Mírovský and Lucie Poláková. 2012a. Analyzing the Most Common Errors in the Discourse Annotation of the Prague Dependency Treebank. In Iris Hendrickx, Sandra Kübler and Kiril Simov (eds.), *Proceedings of the Eleventh International Workshop on Treebanks and Linguistic Theories (TLT11)*. Universidade de Lisboa, Lisbon: Edicoes Colibri, 127–132.

Jínová, Pavlína, Jiří Mírovský and Lucie Poláková. 2012b. Semi-Automatic Annotation of Intra-Sentential Discourse Relations in PDT. In Eva Hajičová, Lucie Poláková and Jiří Mírovský (eds.), *Proceedings of the Workshop on Advances in Discourse Analysis and its Computational Aspects (ADACA) at Coling 2012*. Mumbai: Coling 2012 Organizing Committee, 43–58.

Jínová, Pavlína, Lucie Poláková and Jiří Mírovský. 2014. Sentence Structure and Discourse Structure (Possible Parallels). In Kim Gerdes, Eva Hajičová and Leo Wanner (eds.), *Dependency Linguistics. Recent Advances in Linguistic Theory Using Dependency Structures*. Amsterdam: Benjamins, 53–74.

Joshi, Aravind and Scott Weinstein. 1981. Control of Inference: Role of Some Aspects of Discourse Structure – Centering. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence – Volume 1*. San Francisco: Morgan Kaufmann, 385–387.

Kehler, Andrew. 2002. *Coherence, Reference, and the Theory of Grammar*. Stanford: CSLI Publications.

Kehler, Andrew, Laura Kertz, Hannah Rohde and Jeffrey L. Elman. 2008. Coherence and Coreference Revisited. *Journal of Semantics*, 25(1), 1–44.

Kehler, Andrew and Hannah Rohde. 2013. A Probabilistic Reconciliation of Coherence-driven and Centering-driven Theories of Pronoun Interpretation. *Theoretical Linguistics*, 39(1–2), 1–37.

Komárek, Miroslav, Jan Petr, Jan Kořenský, Anna Jirsová, Naďa Svozilová, Karel Hausenblas, Jan Balhar, Emil Dvořák, Milena Rulfová, Zdeňka Hrušková, Jarmila Panevová, Eva Buráňová, Libuše Kroupová and Oldřich Uličný. 1986. *Mluvnice češtiny. 2. Tvarosloví [Grammar of Czech. 2. Morphology]*. Prague: Academia.

Koo, Terry, Alexander M. Rush, Michael Collins, Tommi Jaakkola and David Sontag. 2010. Dual Decomposition for Parsing with Non-Projective Head Automata. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg: Association for Computational Linguistics, 1288–1298.

Korzen, Iørn and Matthias Buch-Kromann. 2011. Anaphoric Relations in the Copenhagen Dependency Treebanks. In Stephanie Dipper and Heike Zinsmeister (eds.), *Beyond Semantics. Corpus-based Investigations of Pragmatic and Discourse Phenomena*. Bochum: Ruhr-Universität Bochum, 83–98.

Krejdlin, Grigory E. and Ekaterina V. Rachilina. 1981. Denotativnyj status otglagoľnych imen [Denotation of Verbal Nouns]. *Naučno-techničeskaja informacija*, 12, 17–22.

Kuno, Susumu. 1972. Functional Sentence Perspective: A Case Study from Japanese and English. *Linguistic Inquiry*, 3(3), 269–320.

Lambrecht, Knud. 1996. *Information Structure and Sentence Form: Topic, Focus, and the Mental Representations of Discourse Referents*. Cambridge, UK: Cambridge University Press.

Langacker, Ronald. 2008. *Cognitive Grammar: A Basic Introduction*. New York: Oxford University Press.

Lapshinova-Koltunski, Ekaterina and Kerstin Kunz. 2014. Annotating Cohesion for Multilingual Analysis. In Harry Bunt (ed.), *Proceedings 10th Joint ISO–ACL SIGSEM Workshop on Interoperable Semantic Annotation*. Reykjavik: European Language Resources Association, 57–64.

Lee, Heeyoung, Angel X. Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu and Dan Jurafsky. 2013. Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules. *Computational Linguistics*, 39(4), 885–916.

Löbner, Sebastian. 1996. Definite Associative Anaphora. In Simon Botley (ed.), *Approaches to Discourse Anaphora: Proceedings of DAARC96 – Discourse Anaphora and Resolution Colloquium*. Lancaster: UCREL Technical Papers Series, 1–22.

Mann, William C. and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text*, 8(3), 243–281.

Martin, James R. 1992. *English Text: System and Structure*. Amsterdam: Benjamins.

Mathesius, Vilém. 1907. Studie k dějinám anglického slovosledu [A Study on History of English Word Order]. *Věstník České akademie*, 16(1), 261–265.

Mathesius, Vilém. 1929. Zur Satzperspektive im modernen Englisch. *Archiv für das Studium der neueren Sprachen und Literaturen*, 155(29), 202–210.

Mathesius, Vilém. 1939. O tak zvaném aktuálním členění větném [On So-Called Topic–Focus Articulation]. *Slovo a slovesnost*, 5(4), 171–174.

Mel'čuk, Igor A. 1981. Meaning–Text Models: A Recent Trend in Soviet Linguistics. *Annual Review of Anthropology*, 10(1), 27–62.

Mendoza, Imke. 2004. *Nominaldetermination im Polnischen*. Ph.D. thesis, Ludwig-Maximilians-Universität, Institut für Slavische Philologie, München.

Mikulová, Marie. 2011. *Významová reprezentace elipsy [Semantic Representation of Ellipsis]*. Prague: Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics.

Mikulová, Marie, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá and Zdeněk Žabokrtský. 2006. *Annotation on the Tectogrammatical Layer in the Prague Dependency Treebank. Annotation Manual*. Technical report TR-2006-30. Prague: Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics. [cit. 2015_08_30]. Available from <https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/pdf/t-man-en.pdf>.

Mikulová, Marie, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá and Zdeněk Žabokrtský. 2005. *Anotace na tektogramatické rovině Pražského závislostního korpusu. Anotátorská příručka [Annotation on the Tectogrammatical Layer in the Prague Dependency Treebank. Annotator´s Manual]*. Technical report TR-2005-28. Prague: Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics. [cit. 2015_08_30]. Available from <https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/cz/t-layer/pdf/t-man-cz.pdf>.

Mikulová, Marie and Jan Štěpánek. 2010. Ways of Evaluation of the Annotators in Building the Prague Czech-English Dependency Treebank. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner and Daniel Tapias (eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta: European Language Resources Association, 1836–1839.

Miltsakaki, Eleni, Rashmi Prasad, Aravind Joshi and Bonnie Webber. 2004. Annotating Discourse Connectives and Their Arguments. In Adam Meyers (ed.), *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*. Boston: Association for Computational Linguistics, 9–16.

Miltsakaki, Eleni, Livio Robaldo, Alan Lee and Aravind Joshi. 2008. Sense Annotation in the Penn Discourse Treebank. In Alexander Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing*. Berlin/Heidelberg: Springer, 275–286.

Mírovský, Jiří. 2009. *Searching in the Prague Dependency Treebank*. Prague: Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics.

Mírovský, Jiří and Eva Hajičová. 2014. What Can Linguists Learn from Some Simple Statistics on Annotated Treebanks. In Verena Henrich, Erhard Hinrichs, Daniël de Kok, Petya Osenova and Adam Przepiórkowski (eds.), *Proceedings of 13th International Workshop on Treebanks and Linguistic Theories (TLT13)*. Tübingen: University of Tübingen, 279–284.

Mírovský, Jiří, Pavlína Jínová and Lucie Poláková. 2012. Does Tectogrammatics Help the Annotation of Discourse? In Martin Kay and Christian Boitet (eds.), *Proceedings of the 24th International Conference on Computational Linguistics*. Mumbai: Coling 2012 Organizing Committee, 853–862.

Mírovský, Jiří, Pavlína Jínová and Lucie Poláková. 2014. Discourse Relations in the Prague Dependency Treebank 3.0. In Lamia Tounsi and Rafal Rak (eds.), *The 25th International Conference on Computational Linguistics, Proceedings of the Conference System Demonstrations*. Dublin: Dublin City University, 34–38.

Mírovský, Jiří, Lucie Mladová and Šárka Zikánová. 2010. Connective-Based Measuring of the Inter-Annotator Agreement in the Annotation of Discourse in PDT. In Chu-Ren Huang and Dan Jurafsky (eds.), *Proceedings of the 23rd International Conference on Computational Linguistics*. Beijing: Tsinghua University Press, 775–781.

Mírovský, Jiří, Kateřina Rysová, Magdaléna Rysová and Eva Hajičová. 2013. (Pre-)Annotation of Topic–Focus Articulation in Prague Czech-English Dependency Treebank. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. Nagoya: Asian Federation of Natural Language Processing, 55–63.

Miéville, Denis. 1999. Associative Anaphora: An Attempt at Formalisation. *Journal of Pragmatics*, 31(3), 327–337.

Mladová, Lucie. 2008. K problematice vztahu rematizátorů a textových konektorů [On the Relationship of Rhematizers and Discourse Connectives]. *Čeština doma a ve světě*, 16(3–4), 126–133.

Mladová, Lucie, Šárka Zikánová, Zuzanna Bedřichová and Eva Hajičová. 2009. Towards a Discourse Corpus of Czech. In Michaela Mahlberg, Victorina González-Díaz and Catherine Smith (eds.), *Proceedings of the Corpus Linguistics Conference (CL 2009)*. Liverpool: University of Liverpool, 8 p. [cit. 2015_07_22]. Available from <http://ucrel.lancs.ac.uk/publications/cl2009/#papers>.

Mladová, Lucie, Šárka Zikánová and Eva Hajičová. 2008. From Sentence to Discourse: Building an Annotation Scheme for Discourse Based on Prague Dependency Treebank. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis and Daniel Tapias (eds.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech: European Language Resources Association, 2564–2570.

Nedoluzhko, Anna and Jiří Mírovský. 2011. *Annotating Extended Textual Coreference and Bridging Relations in the Prague Dependency Treebank*. Technical report TR-2011-44. Prague: Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics. [cit. 2015_08_30]. Available from <http://ufal.mff.cuni.cz/techrep/tr44.pdf>.

Nedoluzhko, Anna and Jiří Mírovský. 2013. Annotators' Certainty and Disagreements in Coreference and Bridging Annotation in Prague Dependency Treebank. In Eva Hajičová, Kim Gerdes and Leo Wanner (eds.), *Depling 2013: Proceedings of the Second International Conference on Dependency Linguistics*. Prague: Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, 236–243.

Nedoluzhko, Anna, Jiří Mírovský and Michal Novák. 2013. A Coreferentially Annotated Corpus and Anaphora Resolution for Czech. *Computational Linguistics and Intellectual Technologies*, 12(1), 467–475.

Ogrodniczuk, Maciej, Katarzyna Głowińska, Mateusz Kopeć, Agata Savary and Magdalena Zawisławska. 2015. *Coreference in Polish: Annotation, Resolution and Evaluation*. Berlin: Walter De Gruyter.

Paducheva, Elena. 1985. *Vyskazyvanie i ego sootnesennost' s dejstvitel'nost'ju [The Utterance and Its Realization in the Text]*. Moscow: Nauka.

Paggio, Patrizia. 2006. Annotating Information Structure in a Corpus of Spoken Danish. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk and Stelios Piperidis (eds.), *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa: European Language Resources Association, 1606–1609.

Pajas, Petr and Jan Štěpánek. 2008. Recent Advances in a Feature-Rich Framework for Treebank Annotation. In Donia Scott and Hans Uszkoreit (eds.), *Proceedings of the 22nd International Conference on Computational Linguistics*. Manchester: The Coling 2008 Organizing Committee, 673–680.

Pajas, Petr and Jan Štěpánek. 2009. System for Querying Syntactically Annotated Corpora. In Gary Lee and Sabine Schulte im Walde (eds.), *Proceedings of the ACL–IJCNLP 2009 Software Demonstrations*. Suntec: Association for Computational Linguistics, 33–36.

Palek, Bohumil. 1968. *Cross-reference: a Study from Hyper-syntax*. Prague: Charles University in Prague.

Palek, Bohumil. 1988. *Referenční výstavba textu [The Referential Structure of the Text]*. Prague: Charles University in Prague.

Panevová, Jarmila. 1991. Koreference gramatická nebo textová? [Grammatical or Textual Coreference?]. In Wieslaw Banys, Leszek Bednarczuk and Krzysztof Bogacki (eds.), *Études de linguistique romane*. Cracow, 495–506.

Paul, Hermann. 1886. *Prinzipien der Sprachgeschichte*. Halle: Max Niemeyer.

Pešek, Ondřej. 2011. *Argumentativní konektory v současné francouzštině a češtině [Argumentative Connectives in Modern French and Czech]*. České Budějovice: Editio Universitatis Bohemiae Meridionalis.

Poesio, Massimo. 2000. *The GNOME Annotation Scheme Manual*. Technical report. Essex: University of Essex. [cit. 2015_08_30]. Available from <http://cswww.essex.ac.uk/Research/nle/corpora/GNOME/anno_manual_4.htm>.

Poesio, Massimo. 2003. Associative Descriptions and Salience: A Preliminary Investigation. In Robert Dale, Kees van Deemter and Ruslan Mitkov (eds.), *Proceedings of EACL 2003 Workshop on the Computational Treatment of Anaphora*. Budapest: Association for Computational Linguistics, 31–28.

Poesio, Massimo and Ron Artstein. 2008. Anaphoric Annotation in the ARRAU Corpus. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis and Daniel Tapias (eds.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech: European Language Resources Association, 1170–1174.

Poesio, Massimo, Rodolfo Delmonte, Antonella Bristot, Luminita Chiran and Sara Tonelli. In prep. *The VENEX Corpus of Anaphora and Deixis in Spoken and Written Italian*. [cit. 2015_09_30]. Available from <http://cswww.essex.ac.uk/staff/poesio/publications/VENEX04.pdf>.

Poesio, Massimo, Renata Vieira and Simone Teufel. 1997. Resolving Bridging References in Unrestricted Text. In Ruslan Mitkov and Branimir Boguraev (eds.), *Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*. Madrid: Association for Computational Linguistics, 1–6.

Poláková, Lucie. 2015. *Discourse Relations in Czech*. Ph.D. thesis, Charles University in Prague, Prague.

Poláková, Lucie, Pavlína Jínová and Jiří Mírovský. 2012. Interplay of Coreference and Discourse Relations: Discourse Connectives with a Referential Component. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk and Stelios Piperidis (eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul: European Language Resources Association, 146–153.

Poláková, Lucie, Pavlína Jínová and Jiří Mírovský. 2014. Genres in the Prague Discourse Treebank. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard and Joseph Mariani (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik: European Language Resources Association, 1320–1326.

Poláková, Lucie, Pavlína Jínová, Šárka Zikánová, Zuzanna Bedřichová, Jiří Mírovský, Magdaléna Rysová, Jana Zdeňková, Veronika Pavlíková and Eva Hajičová. 2012a. *Manual for Annotation of Discourse Relations in Prague Dependency Treebank*. Technical report TR-2012-47. Prague: Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics. [cit. 2015_08_30]. Available from <http://ufal.mff.cuni.cz/techrep/tr47.pdf>.

Poláková, Lucie, Pavlína Jínová, Šárka Zikánová, Eva Hajičová, Jiří Mírovský, Anna Nedoluzhko, Magdaléna Rysová, Veronika Pavlíková, Jana Zdeňková, Jiří Pergler and Radek Ocelák. 2012b. *Prague Discourse Treebank 1.0*. Data/software. Prague: Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics. [cit. 2015_07_22].

Poláková, Lucie, Jiří Mírovský, Anna Nedoluzhko, Pavlína Jínová, Šárka Zikánová and Eva Hajičová. 2013. Introducing the Prague Discourse Treebank 1.0. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. Nagoya: Asian Federation of Natural Language Processing, 91–99.

Pradhan, Sameer, Alessandro Moschitti, Nianwen Xue, Olga Uryupina and Yuchen Zhang. 2012. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. In *Proceedings of the Joint Conference on EMNLP and CoNLL: Shared Task*. Stroudsburg: Association for Computational Linguistics, 1–40.

Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis and Daniel Tapias (eds.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech: European Language Resources Association, 2961–2968.

Prasad, Rashmi, Aravind Joshi and Bonnie Webber. 2010. Realization of Discourse Relations by Other Means: Alternative Lexicalizations. In Chu-Ren Huang and Dan Jurafsky (eds.), *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Beijing: Tsinghua University Press, 1023–1031.

Prasad, Rashmi, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo and Bonnie Webber. 2007. *The Penn Discourse Treebank 2.0 Annotation Manual*. Technical report IRCS-08-01. Philadelphia: Institute for Research in Cognitive Science, University of Pennsylvania. [cit. 2015_08_30]. Available from <http://www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb-annotation-manual.pdf>.

Prasad, Rashmi, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi and Bonnie Webber. 2006. *The Penn Discourse Treebank 1.0 Annotation Manual*. Technical report IRCS-06-01. Philadelphia: Institute for Research in Cognitive Science, University of Pennsylvania. [cit. 2015_08_30]. Available from <http://www.seas.upenn.edu/~pdtb/papers/pdtb-1.0-annotation-manual.pdf>.

Prasad, Rashmi, Bonnie Webber and Aravind Joshi. 2014. Reflections on the Penn Discourse TreeBank, Comparable Corpora, and Complementary Annotation. *Computational Linguistics*, 40(4), 921–950.

Prasad, Rashmi, Bonnie Webber, Alan Lee and Aravind Joshi. In prep. Discourse Relations in the PDTB 3.0.

Prince, Ellen F. 1981. Toward a Taxonomy of Given–New Information. In Peter Cole (ed.), *Radical Pragmatics*. New York: Academic Press, 223–255.

Putnam, Hilary. 1961. Some Issues in the Theory of Grammar. In Roman Jakobson (ed.), *The Structure of Language and Its Mathematical Aspects, Proceedings of Symposia in Applied Mathematics*. Providence: American Mathematical Society, 25–42.

Recasens, Marta, Eduard Hovy and Maria Antònia Martí. 2010. A Typology of Near-Identity Relations for Coreference (NIDENT). In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner and Daniel Tapias (eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta: European Language Resources Association, 149–156.

Recasens, Marta and Maria Antònia Martí. 2010. AnCora-CO: Coreferentially Annotated Corpora for Spanish and Catalan. *Language Resources and Evaluation*, 44(4), 315–345.

Recasens, Marta, Maria Antònia Martí and Mariona Taulé. 2007. Text as Scene: Discourse Deixis and Bridging Relations. *Procesamiento del lenguaje natural*, 39, 205–212.

Rejzek, Jiří. 2004. *Český etymologický slovník [Czech Etymological Dictionary]*. Prague: Leda.

Russel, Bertrand. 1905. On Denoting. *Mind*, 14(56), 479–493.

Rysová, Kateřina. 2011. The Word Order of Inner Participants in Czech, Considering the Systemic Ordering of Actor and Patient. In Kim Gerdes, Eva Hajičová and Leo Wanner (eds.), *Depling 2011: Proceedings, International Conference on Dependency Linguistics*. Barcelona: Universitat Pompeu Fabra, 183–192.

Rysová, Kateřina. 2014a. *O slovosledu z komunikačního pohledu [On Word Order from the Communicative Point of View]*. Prague: Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics.

Rysová, Kateřina and Jiří Mírovský. 2014a. Valency and Word Order in Czech – A Corpus Probe. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard and Joseph Mariani (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik: European Language Resources Association, 975–980.

Rysová, Magdaléna. 2012. Alternative Lexicalizations of Discourse Connectives in Czech. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk and Stelios Piperidis (eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul: European Language Resources Association, 2800–2807.

Rysová, Magdaléna. 2014b. Verbs of Saying with a Textual Connecting Function in the Prague Discourse Treebank. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard and Joseph Mariani (eds.), *Proceedings of the Ninth International*

Conference on Language Resources and Evaluation (LREC'14). Reykjavik: European Language Resources Association, 930–935.

Rysová, Magdaléna. 2015. *Diskurzní konektory v češtině (Od centra k periferii) [Discourse Connectives in Czech (From Centre to Periphery)]*. Ph.D. thesis, Charles University in Prague, Prague.

Rysová, Magdaléna and Jiří Mírovský. 2014b. Use of Coreference in Automatic Searching for Multiword Discourse Markers in the Prague Dependency Treebank. In Lori Levin and Manfred Stede (eds.), *The 8th Linguistic Annotation Workshop in Conjunction with Coling 2014, Proceedings of the Workshop*. Dublin: Dublin City University, 11–19.

Rysová, Magdaléna and Kateřina Rysová. 2014. The Centre and Periphery of Discourse Connectives. In Wirote Aroonmanakun, Prachya Boonkwan and Thepchai Supnithi (eds.), *Proceedings of Pacific Asia Conference on Language, Information and Computing*. Bangkok: Department of Linguistics, Faculty of Arts, Chulalongkorn University, 452–459.

Rysová, Magdaléna and Kateřina Rysová. 2015. Secondary Connectives in the Prague Dependency Treebank. In Joakim Nivre and Eva Hajičová (eds.), *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*. Uppsala: Uppsala University, 291–299.

Sauper, Christina, Aria Haghighi and Regina Barzilay. 2010. Incorporating Content Structure into Text Analysis Applications. In Hang Li and Lluís Màrquez (eds.), *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Cambridge, MA, USA: Association for Computational Linguistics, 377–387.

Schwarz-Friesel, Monika. 2007. Indirect Anaphora in Text. A Cognitive Account. In Monika Schwarz-Friesel, Manfred Consten and Mareile Knees (eds.), *Anaphors in Text. Cognitive, Formal and Applied Approaches to Anaphoric Reference*. Amsterdam: Benjamins, 3–20.

Ševčíková, Magda and Jiří Mírovský. 2012. Sentence Modality Assignment in the Prague Dependency Treebank. In Petr Sojka, Aleš Horák, Ivan Kopeček and Karel Pala (eds.), *Text, Speech and Dialogue: 15th International Conference, TSD 2012*. Berlin/Heidelberg: Springer, 56–63.

Sgall, Petr. 1967a. Functional Sentence Perspective in a Generative Description of Language. *Prague Studies in Mathematical Linguistics*, 2, 203–225.

Sgall, Petr. 1967b. *Generativní popis jazyka a česká deklinace [Generative Description of Language and Czech Declension]*. Prague: Academia.

Sgall, Petr. 1975. On the Nature of Topic and Focus. In Hakan Ringbom (ed.), *Style and Text (Studies Presented to Nils Erik Enkvist)*. Stockholm: Scriptor, 409–15.

Sgall, Petr. 1979. Towards a Definition of Focus and Topic. *Prague Bulletin of Mathematical Linguistics*, 31, 3–25.

Sgall, Petr, Eva Hajičová and Eva Buráňová. 1980. *Aktuální členění věty v češtině [Topic–Focus Articulation in Czech]*. Prague: Academia.

Sgall, Petr, Eva Hajičová and Eva Benešová. 1973. *Topic, Focus and Generative Semantics*. Kronberg/Taunus: Scriptor.

Sgall, Petr, Eva Hajičová and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Dordrecht: Reidel Publishing Company.

Sgall, Petr, Ladislav Nebeský, Alla Goralčíková and Eva Hajičová. 1969. *A Functional Approach to Syntax in Generative Description of Language*. New York: American Elsevier Publishing Company.

Shmelev, Alexey D. 1996. *Referencial'nyje mechanizmy russkogo jazyka [Referential Mechanisms of Russian]*. Tampere: Slavica Tamperensia.

Spoustová, Drahomíra. 2008. Combining Statistical and Rule-Based Approaches to Morphological Tagging of Czech Texts. *Prague Bulletin of Mathematical Linguistics*, 89, 23–40.

Stede, Manfred. 2004. The Potsdam Commentary Corpus. In Bonnie Webber and Donna Byron (eds.), *Proceedings of the 2004 ACL Workshop on Discourse Annotation*. Stroudsburg: Association for Computational Linguistics, 96–102.

Stede, Manfred and Arne Neumann. 2014. Potsdam Commentary Corpus 2.0: Annotation for Discourse Research. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk and Stelios Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik: European Language Resources Association, 925–929.

Steedman, Mark. 1991. Structure and Intonation. *Language*, 67(2), 260–296.

Štěpánková, Barbora. 2014. *Aktualizátory ve výstavbě textu, zejména z pohledu aktuálního členění [Foculizing Particles in Text Structuring, especially from Topic–Focus Articulation Perspective]*. Prague: Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics.

Stepanov, Jury S. 2004. *Imena, predikaty, predloženija [Nouns, Predicates, Sentences]*. Moscow: Slavica Tamperensia.

Taboada, Maite, Julian Brooke and Manfred Stede. 2009. Genre-Based Paragraph Classification for Sentiment Analysis. In Patrick Healey, Roberto Pieraccini, Donna Byron, Steve Young and Matthew Purver (eds.), *Proceedings of the SIGDIAL 2009 Conference. The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Stroudsburg: Association for Computational Linguistics, 62–70.

Taboada, Maite and Debopam Das. 2013. Annotation upon Annotation. Adding Signalling Information to a Corpus of Discourse Relations. *Dialogue and Discourse*, 4(2), 249–281.

Toldova, Svetlana, Anna Nedoluzhko, Asja Rojtberg, Alina Ladygina, Maria Vasilyeva, Ilja Azerkovich, Matvej Kurzukov, Anastasija Ivanova and Julia Grishina. 2014. RU-EVAL-2014: Evaluating Anaphora and Coreference Resolution for Russian. *Computational Linguistics and Intellectual Technologies*, 13 (20), 681–694.

Václ, Jan. 2015. *Tracing Salience in Documents*. Master's thesis, Charles University in Prague, Prague.

Veselá, Kateřina, Jiří Havelka and Eva Hajičová. 2004. Annotators' Agreement: The Case of Topic–Focus Articulation. In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa and Raquel Silva (eds.), *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon: European Language Resources Association, 2191–2194.

Veselá, Kateřina, Nino Peterek and Eva Hajičová. 2003. Some Observations on Contrastive Topic in Czech Spontaneous Speech. *Prague Bulletin of Mathematical Linguistics*, 79–80, 5–22.

Webber, Bonnie. 2009. Genre Distinctions for Discourse in the Penn TreeBank. In Keh-Yih Su, Jian Su, Janyce Wiebe and Haizhou Li (eds.), *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*. Suntec: Association for Computational Linguistics, 674–682.

Webber, Bonnie, Alistair Knott and Aravind Joshi. 2001. Multiple Discourse Connectives in a Lexicalized Grammar for Discourse. In Harry Bunt, Reinhard Muskens and Elias Thijsse (eds.), *Computing Meeting, Volume 2*. Tilburg: Kluwer Academic Publishers, 229–245.

Wegener, Philipp. 1885. *Untersuchungen über die Grundfragen des Sprachlebens*. Amsterdam: Benjamins.

Weil, Henri. 1844. *Question de grammaire générale: de l'ordre des mots dans les langues anciennes comparées aux langues modernes (thèse française)*. Paris: Joubert. Transl. by Charles W. Super as Weil, Henri. 1887. *The Order of Words in the Ancient Languages Compared with That of the Modern Languages*. Boston: Ginn.

Weil, Henri. 1978. *The Order of Words in the Ancient Languages Compared with That of the Modern Languages*. Amsterdam/Philadelphia: Benjamins. Aldo Scaglione (ed.), edited and reprinted version of Weil (1887), see Weil (1844).

Wolf, Florian and Edward Gibson. 2005. Representing Discourse Coherence: A Corpus-based Study. *Computational Linguistics*, 31(2), 249–287.

Zeman, Daniel, Ondřej Dušek, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zdeněk Žabokrtský and Jan Hajič. 2014. HamleDT: Harmonized Multi-Language Dependency Treebank. *Language Resources and Evaluation*, 48(4), 601–637.

Zikánová, Šárka. 2006. What do the Data in Prague Dependency Treebank Say about Systemic Ordering in Czech? *The Prague Bulletin of Mathematical Linguistics*, 86, 39–46.

Zikánová, Šárka, Lucie Poláková, Pavlína Jínová, Anna Nedoluzhko, Magdaléna Rysová, Jiří Mírovský and Eva Hajičová. 2015. Zachycení výstavby textu v Pražském závislostním korpusu [Annotation of Discourse Phenomena in the Prague Dependency Treebank]. *Slovo a slovesnost*, 76(3), 163–197.

Zikánová, Šárka, Miroslav Týnovský and Jiří Havelka. 2007. Identification of Topic and Focus in Czech: Evaluation of Manual Parallel Annotations. *The Prague Bulletin of Mathematical Linguistics*, 87, 61–70.

Sources

Bejček, Eduard, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Magda Ševčíková, Jan Štěpánek and Šárka Zikánová. 2013. *Prague Dependency Treebank 3.0.* Data/software. Prague: Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics. [cit. 2015_07_22].

Hajič, Jan, Silvie Cinková, Kristýna Čermáková, Lucie Mladová, Anna Nedoluzhko, Petr Pajas, Jiří Semecký, Jana Šindlerová, Josef Toman, Kristýna Tomšů, Matěj Korvas, Magdaléna Rysová, Kateřina Veselovská and Zdeněk Žabokrtský. 2009. *Prague English Dependency Treebank 1.0.* Data/software. Prague: Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics. [cit. 2015_07_22].

Hajič, Jan, Eva Hajičová, Jarmila Panevová, Petr Sgall, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová and Zdeněk Žabokrtský. 2011. *Prague Czech-English Dependency Treebank 2.0.* Data/software. Prague: Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics. [cit. 2015_07_22].

Asher, Nicholas and Alex Lascarides. 1998. Bridging. *Journal of Semantics*, 15(1), 83–113.

Carroll, Lewis. 1865. *Alice in Wonderland.* London: Macmillan.

Clark, Herbert H. 1975. Bridging. In David Waltz (ed.), *Theoretical Issues in Natural Language Processing.* New York: Association for Computing Machinery, 169–174.

Krejdlin, Grigory E. and Ekaterina V. Rachilina. 1981. Denotativnyj status otglagoľnych imen [Denotation of Verbal Nouns]. *Naučno-techničeskaja informacija,* 12, 17–22.

Mikulová, Marie, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá and Zdeněk Žabokrtský. 2005. *Anotace na tektogramatické rovině Pražského závislostního korpusu. Anotátorská příručka [Annotation on the Tectogrammatical Layer in the Prague Dependency Treebank. Annotator's Manual].* Technical report TR-2005-28. Prague: Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics. [cit. 2015_08_30]. Available from <https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/cz/t-layer/pdf/t-man-cz.pdf>.

Perspective Digest. [cit. 2015_08_30]. Available from <http://www.perspectivedigest.org>.

Škvorecký, Josef. 1986. *Dvorak in Love. A light-hearted dream.* Transl. from the Czech original *Scherzo capriccioso* by Paul Wilson. Toronto: Lester & Orpen Dennys.

Škvorecký, Josef. 1991. *Scherzo capriccioso: Veselý sen o Dvořákovi.* Prague: Odeon.

Subject Index

adverb, 151, 160
alternative lexicalization of a discourse connective, 15, 23, 166
AltLex, *see* alternative lexicalization of a discourse connective
anaphor, 31
anaphora, 31
  without coreference, 59
anaphoric reference, 171
annotation scenario, 83
antecedent, 31, 56
apposition, 121, 122
argument, *see* discourse argument
attachment score, 90
attribute
  afun, 84
  coref_special, 106
  deepord, 86
  functor, 84
  tfa, 84, 86
attribution, 15, 23, 186, 188
  in the PDTB, 181

background, 63
Biomedical Discourse Relation Bank, 182
bridging anaphora, 4, 6, 51, 86, 88, 92, 106

caption, 26, 27
cataphora, 31, 191, 195
clause, 17, 169
  semi-clause, 169
  structure, 169
Cohen's κ, 90
coherence, 3, 13, 27, 32, 180, 182
  disruption, 180, 186, 195
  reader's expectation, 190
  weak, 180
cohesion, 3, 51, 179
cohyponymy, 191, 203
comment (RST term), 180
communicative dynamism, 65, 66, 70, 78, 86
conjunction, 150, 159, 176
  coordinate, 156, 160
  subordinate, 156, 160, 161
connective, *see* discourse connective
context, 67
contextual boundness, 66–68, 76, 86, 88, 91, 197
  contrastive, 68
contrast, 58, 66, 68
control, 35
coordinate structure, 118, 129, 143
coordination resolution, 134
coreference, 6, 29, 86, 92, 105, 219
  grammatical, 33
  textual, 33, 88
coreference chain, 32, 221
  length, 222
coreference link, 4, 221
coreference resolution, 33
coulisse, 207
Czech language, 9

deep structure, *see* layer of annotation, tectogrammatical
degree of connectivity, 161
degree of salience, 216, 224
deixis, 204
dependency tree, 84, 118
discourse, 3, 190
discourse analysis, 166

discourse annotation, automatic, 121
  connective detection, 138, 139
  scope of argument, 132, 134
  type of relation, 130
discourse annotation, manual, 121
  connective detection, 136, 137
  scope of argument, 132
discourse argument, 13, 14, 17, 18, 165
  incomplete, 185
discourse connective, 14–16, 18, 23, 24, 166
  aby [*in order to*], 151
  což [*which*], 139, 152
  s tím, že [*along with*], 124
  totiž [*actually, in fact*], 155
  totiž [*in fact*], 145
  frequency, 157
  inflection, 151
  inter-sentential use, 14, 156, 160
  intra-sentential use, 156, 160
  origin, 153
  part-of-speech classification, 150, 159, 160, 165
  placement, 155
  primary, 16, 167, 168, 176
  secondary, 16, 23, 166, 167, 176
discourse entity, 29
discourse relation, 4, 6, 13, 14, 22, 27, 86, 87, 92, 104
  concession, 145
  conjunction, 142
  explication, 19, 145
  opposition, 142
  entity-based, 15, 23
  explicit, 23, 165
  implicit, 15, 23, 165, 184
  inter-sentential, 14, 24, 140
  intra-sentential, 14, 15, 24, 25, 140
  semantic class, 142
  semantic type, 14, 18, 20
  unsignaled, 180
disjunct, 169

effective subtree, 132
ellipsis, 135, 191, 193
ellipsis resolution, 135

endophoric reference, 30
EntRel, *see* discourse relation, entity-based
exophoric reference, 30

F1-measure, 90
  chain-based, 92
  connective-based, 92
focus, 65, 66, 72
  focus proper, 73, 79
Functional Generative Description, 14, 66, 83
functional sentence perspective, *see* topic–focus articulation
functor, 14, 19, 25, 121, 129, 130
  adversative relation, 126
  conjunction, 127
  expression referring to preceding context, 136

general topic (hypertopic), 203
generic reference, 30
genre of document, 27, 28, 88, 94
genre rules, 191, 192
grammatemes, 84

heading, 27
hierarchy of discourse senses (PDTB), 15, 19
hyperonymy, 53, 191
hyponymy, 53, 191

identity of referents, 30
information (in terms of attribution), 188
information structure, *see* topic–focus articulation
inter-annotator agreement, 89
intonation centre, 79

layer of annotation
  analytical, 83, 84, 174
  morphological, 83, 84
  tectogrammatical, 4, 14, 23, 83, 84, 117
  word, 84
lexical bundle, 174
list structure, 26, 191

machine learning, 229
meronymy, 55
metatext, 27
minimality principle, 17
multiword discourse phrase, *see* discourse connective, secondary
multiword expression, 88, 91

near-identity, 31
NEGRA, 90
no relation, 179, 180
  in the PDTB, 15, 181
NoRel, *see* no relation

object–function, 58

part–whole, 55
particle, 151, 166
  rhematizer, 64, 137, 156
Penn Discourse Treebank, 14, 19, 166, 181
PML-Tree Query, 97
  discourse relation, 104
  member node, 104
  negation, 101
  non-dependency relation, 103
  output filter, 107, 108
  search server, 114
postcedent, 31
Prague Czech-English Dependency Treebank, 91
Prague Dependency Treebank, 7, 13, 27, 83, 89, 166, 219
  getting the data, 111
precision, 90
predicate, 65
prepositional phrase, 171
psychological predicate, 65
psychological subject, 65
punctuation mark, 16, 149, 157

quasi-control, 35
question test, 74
question–answer relation, 189

recall, 90
reference, 30
  anaphoric, 199
  endophoric, 36
  exophoric, 43
  generic, 30
  specific, 30
relative clause, 123
rhematizer, *see* particle, rhematizer
rheme, 63
Rhetorical Structure Theory, 179
RST Discourse Treebank, 180

salience, 4, 6, 216
  graph, 224
  leap height, 226
salience algorithm, 222, 223
searching in the PDT, 97, 114
searching in treebanks, 97
sentence, 17
sentence element, 169
sentence modality, 91
set–subset, 56
situational context, 67
solutionhood relation, 189
source of information (in terms of attribution), 188
specific reference, 30
stock of shared knowledge, 4, 216
subordinate structure, 118, 143
summary (RST term), 181
systemic ordering, 71

test with negation, 75
text, *see* discourse
text-organizing devices, 186, 187
textuality, 179, 190
thematic continuity, 191
theme, 63
topic, 65, 66
  proper, 72
  sentence without topic, 73
topic–focus articulation, 4, 5, 63, 65, 67, 84, 91, 105, 223
topic-shift (RST term), 181

tree
  analytical, 229
  tectogrammatical, 229
Tree Editor TrEd, 112
  extension, 114

valency, 84

verbal phrase, 171

word order
  deep, 66, 70, 79
  free, 79
  surface, 70
world knowledge, 191, 204

Name Index

Adamec, Přemysl, 33, 39
Artstein, Ron, 33
Arutunova, Natalia D., 41
Asher, Nicholas, 13, 15, 17, 51, 53

Balhar, Jan, 150
Bärenfänger, Maja, 54, 55
Barzilay, Regina, 218
Bauer, Jaroslav, 151, 153
Baumann, Stefan, 76
de Beaugrande, Robert-Alain, 3, 179
Bedřichová, Zuzanna, 13, 17, 19, 126, 151
Bejček, Eduard, 27, 83, 88, 91, 211
Bělič, Jaromír, 159
Běličová, Helena, 124
Bémová, Alevtina, 19, 34–38, 90, 122, 139, 228, 229
Benešová, Eva, 65, 66
Biber, Douglas, 174
Bohnet, Bernd, 76
Brants, Thorsten, 90, 91
Brenier, Jason, 76
Brinckmann, Caren, 76
Bristot, Antonella, 33, 54
Brooke, Julian, 27
Buch-Kromann, Matthias, 54
Buráňová, Eva, 65, 71, 150, 207
Burga, Alicia, 76

Calhoun, Sasha, 76
Carlson, Greg, 30, 33
Carlson, Lynn, 117, 181
Carnap, Rudolf, 32
Čejka, Mirek, 124
Čermák, František, 173
Červená, Vlasta, 159
Chafe, Wallace L., 217
Chambers, Nathanael, 33
Chang, Angel X., 33
Charolles, Michel, 51
Cheng, Hua, 33
Chernejko, Ludmila O., 41
Chiran, Luminita, 33, 54
Chloupek, Jan, 124
Chomsky, Noam, 7, 75
Churavý, Miloslav, 159
Clark, Herbert H., 51–54, 62, 217
De Clercq, Orphée, 94
Cohen, Jacob, 90
Conrad, Susan, 174

Daneš, František, 52, 63, 123, 124, 159, 198
Das, Debopam, 14, 180, 187, 189
Delmonte, Rodolfo, 33, 54
Dinesh, Nikhil, 14, 15, 17, 31, 93, 117, 132, 142, 179, 181, 233
Dipper, Stefanie, 76
Dressler, Wolfgang, 3, 179
Dušek, Ondřej, 97
Dvořák, Emil, 124, 150

Elman, Jeffrey L., 32

Filipec, Josef, 159
Firbas, Jan, 65, 70
Forbes-Riley, Katherine, 174
Frege, Gottlob, 32, 41
Fried, Mirjam, 123

von der Gabelentz, Georg, 65
Gardent, Claire, 54, 55
Gebauer, Jan, 154
Gibson, Edward, 13
Głowińska, Katarzyna, 31, 33

Goecke, Daniela, 54, 55
Goralčíková, Alla, 7, 83
Götze, Michael, 76
Grepl, Miroslav, 124
Grosz, Barbara J., 3, 218
Gundel, Jeanette K., 217

Haghighi, Aria, 218
Hajič, Jan, 19, 27, 34–38, 83, 90, 97, 122, 139, 159, 228, 229
Hajičová, Eva, 3, 7, 13, 17, 19, 23, 27, 33–38, 64–67, 70–72, 75, 76, 83, 91, 92, 95, 118, 122, 126, 139, 184, 207, 218, 223, 224, 228, 229
Halliday, Michael Alexander Kirkwood, 3, 31, 32, 179, 212, 217
Hana, Jiří, 97, 159
Hanová, Hana, 159
Hansen-Schirra, Silvia, 76
Hasan, Ruqaiya, 3, 31, 32, 179, 212
Hausenblas, Karel, 124, 150, 159
Havelka, Jiří, 19, 34–38, 72, 83, 91, 122, 139, 224, 228, 229
Haviland, Susan E., 217
Havlová, Františka, 159
Havránek, Bohuslav, 159
Hedberg, Nancy, 217
Helcl, Miloš, 159
Hendrickx, Iris, 94
Henschel, Renate, 33
Hilbert, Mirco, 54, 55
Hladká, Barbora, 83, 90, 159, 218, 224
Hlaváčková, Zdeňka, 159
Hlaváčová, Jaroslava, 159
Hlavsa, Zdeněk, 30, 39, 124
Hobbs, Jerry R., 13, 32, 190
Hoffmannová, Jana, 3, 124
Hoste, Veronique, 94
Hovorková, Alena, 159
Howarth, Peter, 173
Hrbáček, Josef, 31, 124, 150
Hrušková, Zdeňka, 150

Janský, Ladislav, 159
Jedlička, Alois, 159
Jeřábek, Emil, 159
Jínová, Pavlína, 13, 17, 22, 23, 25, 27, 83, 86, 88, 92, 94, 117, 126, 154, 160, 184
Jirsová, Anna, 150
Joshi, Aravind, 13–15, 17, 19, 31, 93, 117, 132, 142, 166, 167, 173, 174, 179, 181, 182, 218, 233
Jurafsky, Dan, 33

Karlík, Petr, 124
Kehler, Andrew, 13, 32, 190, 212
Kertz, Laura, 32
Kettnerová, Václava, 27, 83
Knott, Alistair, 14
Kolářová, Veronika, 19, 27, 34–38, 83, 122, 139, 228, 229
Komárek, Miroslav, 150
Kopeć, Mateusz, 31, 33
Kořenský, Jan, 150
Korzen, Iørn, 54
Kow, Eric, 54, 55
Kozlová, Karla, 159
Kratochvílová, Helena, 159
Krejdlin, Grigory E., 42
Kroupová, Libuše, 150, 159
Kruijff, Geert-Jan M., 76
Kruijff-Korbayová, Ivana, 76
Kučová, Lucie, 19, 34–38, 122, 139, 218, 224, 228, 229
Kuno, Susumu, 217
Kunz, Kerstin, 33, 212

Lambrecht, Knud, 66
Langacker, Ronald, 29
Lapata, Mirella, 218
Lapshinova-Koltunski, Ekaterina, 33, 212
Lascarides, Alex, 13, 51, 53
Lee, Alan, 14, 15, 17, 19, 31, 93, 117, 132, 142, 179, 181, 233
Lee, Heeyoung, 33
Löbner, Sebastian, 51, 53
Lopatková, Markéta, 19, 34–38, 122, 139, 228, 229
Lüngen, Harald, 54, 55

Machač, Jaroslav, 159
Macháčková, Eva, 124
Mahuelian, Helene, 54, 55
Mann, William C., 13, 179
Marcu, Daniel, 117, 181
Mareček, David, 97
Marešová, Hana, 159
Martí, Maria Antònia, 33, 54, 55
Martin, James R., 13
Mathesius, Vilém, 65
Mejstřík, Vladimír, 159
Mel'čuk, Igor A., 66
Mellish, Chris, 33
Mendoza, Imke, 33, 42
Michálek, Emanuel, 159
Miéville, Denis, 51
Mikulová, Marie, 19, 27, 34–38, 45, 83, 91, 122, 135, 139, 228, 229
Miltsakaki, Eleni, 13–15, 17, 19, 31, 93, 117, 132, 142, 179, 181, 233
Mírovský, Jiří, 13, 17, 22, 23, 25, 27, 55, 70, 76, 83, 86, 88, 91, 92, 94, 95, 97, 117, 126, 154, 160, 175, 184, 203, 229
Moschitti, Alessandro, 93
Müllerová, Olga, 124

Nebeský, Ladislav, 7, 83
Nedoluzhko, Anna, 13, 23, 27, 55, 83, 92, 184, 203, 229
Nekvapil, Jiří, 124
Neumann, Arne, 76, 150
Neumann, Stella, 76
Nissim, Malvina, 76
Novák, Michal, 229
Novotný, Jiří, 124

Ocelák, Radek, 27, 83
Ogrodniczuk, Maciej, 31, 33
Okurowski, Mary Ellen, 117
Oliva, Karel, 33, 34

Paducheva, Elena, 29, 30, 33
Paggio, Patrizia, 76
Pajas, Petr, 19, 34–38, 83, 86, 91, 97, 112, 122, 139, 228, 229
Palek, Bohumil, 30, 31, 39, 124
Panevová, Jarmila, 3, 7, 19, 27, 33–38, 66, 72, 83, 90, 118, 122, 139, 150, 223, 224, 228, 229
Papírníková, Blanka, 159
Partee, Barbara, 66, 67, 70, 72
Paul, Hermann, 65
Pavlíková, Veronika, 13, 17, 27, 83, 126
Pecina, Pavel, 211
Peirsman, Yves, 33
Pelletier, Francis Jeffry, 30, 33
Pergler, Jiří, 27, 83
Pešek, Ondřej, 150, 151
Peterek, Nino, 66
Petr, Jan, 150
Piťha, Petr, 124
Poesio, Massimo, 33, 51, 53–55
Pokorná, Eva, 159
Poláková (Mladová), Lucie, 13, 17, 19, 21–23, 25, 27, 83, 86, 88, 92, 94, 117, 126, 137, 154, 160, 184
Popel, Martin, 97
Popelka, Jan, 83
Poštolková, Běla, 159
Pradhan, Sameer, 93
Prasad, Rashmi, 13–15, 17, 31, 93, 117, 132, 142, 166, 167, 173, 174, 179, 181, 182, 233
Prince, Ellen F., 51, 217
Prouzová, Hana, 124
Putnam, Hilary, 7

Rachilina, Ekaterina V., 42
Ramasamy, Loganathan, 97
Razímová, Magda, 19, 34–38, 122, 139, 228, 229
Recasens, Marta, 33, 54, 55
Rejzek, Jiří, 154
Robaldo, Livio, 14, 15, 17, 19, 31, 93, 117, 132, 142, 233
Rohde, Hannah, 32, 212
Roudný, Miroslav, 159
Rulfová, Milena, 124, 150
Rulíková, Blažena, 124
Russel, Bertrand, 32
Rysová, Kateřina, 16, 70, 71, 76, 163, 166, 167, 173

Rysová, Magdaléna, 13, 16, 17, 27, 76, 83, 126, 154, 163, 166, 167, 173, 175, 176

Sauper, Christina, 218
Savary, Agata, 31, 33
Schwarz-Friesel, Monika, 51
Ševčíková, Magda, 27, 83, 91
Sgall, Petr, 3, 7, 19, 33–38, 65–67, 70–72, 83, 118, 122, 139, 207, 223, 224, 228, 229
Shmelev, Alexey D., 33
Sidner, Candace L., 3
Skopeteas, Stavros, 76
Smejkalová, Lenka, 83
Sochová, Zdeňka, 159
Šoltys, Otakar, 124
Spoustová, Drahomíra, 229
Stede, Manfred, 27, 76, 117, 150
Steedman, Mark, 66, 76
Steiner, Erich, 76
Štěpánek, Jan, 19, 27, 34–38, 83, 86, 91, 97, 112, 122, 139, 228, 229
Štěpánková, Barbora, 64, 137
Stepanov, Jury S., 41
Štindlová, Jitka, 159
Straňák, Pavel, 83, 88, 91, 211
Stührenberg, Maik, 54, 55
Surdeanu, Mihai, 33
Svozilová, Naďa, 150, 159

Taboada, Maite, 14, 27, 180, 187, 189
Taulé, Mariona, 54, 55
Teich, Elke, 76
Teufel, Simone, 51, 53
Thompson, Sandra A., 13, 179
Toman, Josef, 83
Tonelli, Sara, 33, 54
Týnovský, Miroslav, 72

Uhlířová, Ludmila, 124
Uličný, Oldřich, 150
Urešová, Zdeňka, 19, 34–38, 83, 122, 139, 228, 229
Uryupina, Olga, 93
Uszkoreit, Hans, 76

Vachek, Josef, 159
Václ, Jan, 229
Váhala, František, 159
Veselá, Kateřina, 19, 34–38, 66, 91, 122, 139, 224, 228, 229
Vieira, Renata, 51, 53
Vodrážková, Eva, 159
Vrbová, Jarka, 218, 224

Wanner, Leo, 76
Webber, Bonnie, 13–15, 17, 27, 31, 93, 117, 132, 142, 166, 167, 173, 174, 179, 181, 182, 233
Wegener, Philipp, 65
Weil, Henri, 63, 65
Weinstein, Scott, 218
Wolf, Florian, 13

Xue, Nianwen, 93

Žabokrtský, Zdeněk, 19, 34–38, 83, 97, 122, 139, 228, 229
Zacharski, Ron, 217
Zawisławska, Magdalena, 31, 33
Žaža, Stanislav, 124
Zdeňková, Jana, 13, 17, 27, 83, 126
Zeman, Daniel, 97, 159
Zhang, Yuchen, 93
Zikánová, Šárka, 13, 17, 19, 23, 27, 71, 72, 83, 92, 126, 184
Zima, Jaroslav, 159