Secondary Connectives in the Prague Dependency Treebank

Magdaléna Rysová
Charles University in Prague
Faculty of Arts
Institute of Czech Language
and Theory of Communication
Czech Republic
magdalena.rysova@post.cz

Kateřina Rysová
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Czech Republic
katerina.rysova@post.cz

Abstract

The paper introduces a new annotation of discourse relations in the Prague Dependency Treebank (PDT), i.e. the annotation of the so called secondary connectives (mainly multiword phrases like *the condition is, that is the reason why, to conclude, this means* etc.). Firstly, the paper concentrates on theoretical introduction of these expressions (mainly with respect to primary connectives like *and, but, or, too* etc.) and tries to contribute to the description and definition of discourse connectives in general (both primary and secondary). Secondly, the paper demonstrates possibilities of annotations of secondary connectives in large corpora (like PDT). The paper describes general annotation principles for secondary connectives used in PDT for Czech and compares the results of this annotation with annotation of primary connectives in PDT. In this respect, the main aim of the paper is to introduce a new type of discourse annotation that could be adopted also by other languages.

1 Introduction

In the paper, we introduce a new annotation of discourse relations in the Prague Dependency Treebank (PDT) enriched by the so called secondary connectives (i.e. especially by the multiword phrases like *hlavním důvodem je* "the main reason is", *závěr zní* "the conclusion is", *to kontrastuje s tím* "this contrasts with" etc.).

We present how it is possible to annotate such variable (i.e. inflectional and modifiable) structures on big data according to general annotation principles. We believe that our methods may be used also for other languages to enrich the discourse annotations of similar corpora.

2 Theoretical Background – Discourse Connectives in General Overview

Many theoretical approaches of discourse analysis (see projects like Penn Discourse Treebank – Prasad et al., 2008 or Potsdam Commentary Corpus – Stede and Neumann, 2014) are based on detection and annotation of discourse connectives in texts. However, there is not a general agreement on definition as well as terminology concerning these expressions (called besides discourse connectives also pragmatic connectives – van Dijk, 1979, discourse particles – Fischer, 2006 etc.). In this paper, we use the term discourse connectives following the Prague tradition.

Very generally, discourse connectives may be defined as language expressions signaling discourse relations within a text. Most of the authors would agree on typical examples like *and, but, or, when, so, because, yet* etc., i.e. on the central or most frequent discourse connectives. However, the authors differ in dealing with less typical examples like *for this reason, this follows* etc., i.e. in (mostly) multiword phrases allowing variation and inflection (impossibility of inflection is, e.g., one of the criteria used for delimitation of connectives in Potsdam Commentary Corpus – Stede and Neumann, 2014).

From part-of-speech perspective, some authors define discourse connectives as subordinating and coordinating conjunctions, prepositional phrases and adverbs (e.g. Prasad et al., 2008, 2010, 2014; Fraser, 1999), others (like Hansen, 1998; Aijmer, 2002; Schiffrin, 1987) add also particles and nominal phrases.

In this paper, we would like to contribute to this discussion on discourse connectives, to present our definition used in PDT and to bring a

Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015), pages 291–299,
Uppsala, Sweden, August 24–26 2015.

new division of discourse connectives based on a large corpus study.

2.1 Delimitation of Connectives in the Prague Dependency Treebank

During the annotation of authentic Czech texts from PDT, we have met many different possibilities of signaling discourse relations – from one-word, frozen conjunctions like *a* "and" or *ale* "but" to multiword phrases like *stručně řečeno* "simply speaking", *vzhledem k této situaci* "considering this situation", *díky této zkušenosti* "thanks to this experience" etc. All of these expressions somehow contribute to the structuring of discourse, but we felt a need to differentiate such wide group of expressions into subgroups taking into account mainly two aspects: **i) semantically, the suitability of the given expression (in its connective meaning[1]) for different contexts**, **ii) grammatically, the phase of grammaticalization of the given expression**.

i) Semantic delimitation of connectives

The suitability for different contexts divides the expressions into two groups. The first contains expressions that are (in their connective meaning) appropriate for many different contexts, the second includes expressions that are context dependent – see Examples 1, 2 and 3:

(1) *Celý den pršelo.* ***Proto*** *nepůjdu na výlet.*
"It was raining the whole day. **Therefore**, I will not go for a trip."

(2) *Chce se stát slavnou herečkou.* ***Kvůli tomu*** *udělá cokoli.*
"She wants to be a famous actress. **Because of this**, she is able to do anything."

(3) *Ředitel firmy uzavřel řadu podezřelých obchodů.* ***Kvůli této činnosti*** *byl vyšetřován policií.*
"Director of the company has entered into a series of suspicious transactions. **Because of this activity**, he was investigated by the police."

In Examples 1, 2 and 3, there are three expressions signaling a discourse relation of reason and result[2]: *proto* "therefore", *kvůli tomu* "because of this"[3] and *kvůli této činnosti* "because of this activity". However, only the first two are suitable also for the other given contexts (i.e. we may say, e.g., ***Therefore*** */* ***Because of this****, he was investigated by the police.* but not *It was raining the whole day.* ****Because of this activity****, I will not go for a trip*.).

In this respect, we consider *proto* "therefore" and *kvůli tomu* "because of this" suitable as connecting expressions for more contexts (i.e. more "universal") than the expression *kvůli této činnosti* "because of this activity". Generally, we call this suitability **a universality principle** according to which we define discourse connectives. In other words, the expressions like *proto* "therefore" and *kvůli tomu* "because of this" are discourse connectives in our approach, whereas expressions like *kvůli této činnosti* "because of this activity" are not, as they signal discourse relations only in a limited set of contexts (these expressions have of course also the compositional function in the text, but – unlike discourse connectives – they are very far from possible grammaticalization). We call these expressions (like *kvůli této činnosti* "because of this activity") **free connecting phrases**.

ii) Grammatical delimitation of connectives (primary vs. secondary connectives)

Within discourse connectives, we distinguish two categories (mainly in terms of grammaticalization) – primary connectives and secondary connectives (as in M. Rysová and K. Rysová, 2014).

Primary connectives are mainly grammatical (or functional) words whose primary function is to connect two units of a text (they mostly belong to conjunctions and structuring particles[4]). Thus they do not have a role of

[1] We are aware that expressions like *and*, *for*, *on the other hand* etc. have also other (non-connective) meanings. However, these other meanings are not in our interest – we evaluate only expressions in their connective function.

[2] The relation of "reason and result" is in PDT delimited as a causal relation in broader sense (i.e. including both "cause" and "consequence"). The terminology of reason and result was adopted from PDTB (see Prasad et al., 2008).

[3] We understand the whole structure *because of this* as a secondary connective, as **because of* itself is an ungrammatical structure and needs to combine with an anaphoric expression to gain a discourse connecting function. At the same time, there are some present-day primary connectives that historically arose from similar combination of a preposition and demonstrative pronoun (e.g. Czech connective *proto* "therefore" from the preposition *pro* "for" and demonstrative pronoun *to* "this").

[4] We define conjunctions (following the traditional Czech grammar) as grammatical words with primary connecting function (like *ale* "but", *nebo* "or", *a* "and" etc.), structuring

sentence elements and in this sense, they do not affect the sentence syntax. Primary connectives are mostly one-word, lexically frozen expressions. Examples of primary connectives are *ale* "but", *a* "and", *zatímco* "whereas", *protože* "because", *když* "when", *nebo* "or" etc.

Secondary connectives are mainly multiword structures functioning as connectives only in certain collocations. Most of them have a key word signaling given type of discourse relation (the cores may be nouns like *condition, reason, difference* etc., verbs like *to mean, to explain, to cause* etc., prepositions like *due to, because of, despite* etc.). Secondary connectives contain (in contrast to primary) some lexical word or words and have a role of sentence elements (*z tohoto důvodu* "for this reason"), sentence modifiers (*obecně řečeno* "generally speaking") or they may form a separate sentence (*Důvod je jednoduchý.* "The reason is simple."). Secondary connectives are not yet grammaticalized, although they exhibit several features typical for the process of grammaticalization (e.g. weakening of singular and plural distinction, gradual loss of the individual lexical meaning and gaining the primary connecting function as a whole structure etc.). Examples of secondary connectives are *podmínkou je* "the condition is", *to znamená* "this means", *to je důvod, proč* "this is the reason why", *kvůli tomu* "because of this", *z těchto důvodů* "for these reasons" etc.

The main difference between primary and secondary connectives thus lies in grammaticalization – i.e. primary connectives are grammaticalized expressions (although sometimes the grammaticalization is not fully completed, which causes discrepancy among certain parts of speech, especially conjunctions, adverbs and particles). From diachronic point of view, primary connectives arose from other parts of speech and very often from combination of several words and gradually became grammaticalized (e.g. English present-day primary connective *because* arose from *bi cause* "by cause", originally a phrase often followed by a subordinate *that*-clause; it is used as one word probably from around 1400 /see Harper, 2001/).

particles as grammatical words expressing a relation of a speaker to the structure of a text (like *jen* "only", *také* "too" etc.).

3 Discourse Annotation in the Prague Dependency Treebank

The annotation of secondary connectives was carried out on the data of the Prague Dependency Treebank (PDT). PDT contains almost 50 thousand of sentences from the Czech newspaper texts. The advantage of this corpus is that it is annotated on more language levels at once – it contains annotation on morphological, syntactical and syntactico-semantic layers, as well as the annotation of discourse phenomena (i.e. coreference and discourse relations).

Discourse relations have been annotated in two phases – firstly expressed by primary connectives, secondly by secondary connectives.

3.1 Annotation of Primary Connectives in the Prague Dependency Treebank

The annotation of primary connectives has been finished in 2012. The annotation has been carried out on the data of the Prague Dependency Treebank 2.5 (Bejček et al., 2012) and has been published as the Prague Discourse Treebank 1.0 (see Poláková et al., 2012). The annotation follows the Penn Discourse Treebank style (Prasad et al., 2008, 2014), i.e. discourse relations (both inter- and intra-sentential) are annotated between two pieces of a text called discourse arguments (defined as abstract objects according to Asher, 1993). The annotation was limited only to such primary connectives that expressed discourse relations between two verbal arguments containing predication (e.g. clauses, sentences or whole paragraphs). The annotated relation was then assigned one semantic type out of 23 types of relations.[5]

In this phase of annotation, the annotators were also asked to mark all candidates to secondary connectives. Their notes then served as a basis for creating a list of such structures used in the second phase of annotation.

3.2 Annotation of Secondary Connectives in the Prague Dependency Treebank

In the next phase, the first discourse annotation in the Prague Dependency Treebank has been extended by secondary connectives. It contains

[5]Concession, condition, confrontation, conjunction, conjunctive alternative, correction, disjunctive alternative, equivalence, exemplification, explication, pragmatic condition, pragmatic contrast, pragmatic reason, generalization, gradation, opposition, asynchronous, purpose, reason and result, restrictive opposition, specification, synchronous, other.

annotation of both inter- and intra-sentential discourse relations.

The annotation of secondary connectives in PDT was based on the list of potential secondary connectives collected during the first discourse annotation in 2012. All the key words of the collected candidates (like *důvod* "reason", *podmínka* "condition", *znamenat* "to mean" etc.) have been automatically detected in the whole PDT data and then manually sorted (as not all tokens of e.g. the word *podmínka* "condition" have a function of secondary connective) and annotated by human annotators (see Rysová and Mírovský, 2014).

The secondary connectives were annotated on the whole PDT data (i.e. almost 50 thousand of sentences).

Besides the secondary connectives, the new annotation includes also the free connecting phrases (see Section 2.1), as their annotation on big data may allow us to study discourse connectives in deeper and contrastive context. For example, we may see the ratio of universal and non-universal phrases in PDT from which we may learn whether the multiword connecting phrases have a tendency to gradually loosen the bonds to the concrete contexts and to stabilize on one, context independent form. In other words, we may learn how far from primary connectives the majority of multiword structures lies.

A significant difference between the annotations of primary and secondary connectives is that unlike the first annotation in 2012, the extended annotation of secondary connectives contains discourse relations between **both verbal and nominal arguments** (as said above, the annotation of primary connectives concentrated only on arguments expressed by verbal propositions or clauses) – see Example 4:

(4) *Koncert nezačal včas. Důvodem byl pozdní příchod houslisty.*
"The concert has not begun on time. The reason was the late arrival of the violinist."
(= because the violinist has arrived late)

In Example 4, there is a discourse relation of reason and result expressed by the secondary connective *the reason was* between two discourse arguments – the first is represented by the whole clause (*The concert has not begun on time.*), the second argument is nominal (i.e. expressed by the nominal phrase *the late arrival of the violinist*). In this case, the secondary connective cannot be replaced by the primary one – we cannot say something like **protože pozdní příchod houslisty* "*because the late arrival of the violinist". We may see that secondary connectives are not yet fully grammaticalized, which means that they may have a function of various sentence elements, including (among others) subject (*the reason*) and predicate (*was*). Therefore, some of the secondary connectives may be followed by the nominalized discourse argument.

We think that the difference between arguments expressed by a verbal or nominal phrase is purely syntactic (*the late arrival of the violinist* vs. *the violinist has arrived late*). Semantically, the meaning remains almost the same. For this reason, we have annotated all discourse arguments according to their semantics (not syntactic representation)[6].

4 Results and Evaluation

In this part of the paper, we present the main results and characteristics of secondary discourse connectives gained from the annotation in PDT with respect to their comparison with primary connectives.

4.1 Evaluation of Annotations – Inter-Annotator Agreement

The inter-annotator (I-A) agreement of secondary connectives annotation was measured on 500 sentences annotated (simultaneously) by two human annotators.[7] We have focused on two main aspects of their annotation: 1. the overall agreement on existence of the discourse relation (i.e. to which extent the annotators agreed on the fact that there is a discourse relation in the given place of a text expressed by a secondary connective); 2. the agreement on semantic types of discourse relations expressed by secondary connectives (like condition, concession etc.). At the same time, we have compared the results of the inter-annotator agreement of secondary connectives with the primary connectives (Poláková et al., 2013) – see Table 1.[8]

Table 1 demonstrates that the I-A agreement is for primary and secondary connectives comparable. The I-A agreement on the existence

[6] However, we have marked them (technically) differently for easier analysis of final results.

[7] Many thanks to Jiří Mírovský for his kind measuring of the I-A agreement.

[8] The existence of discourse relation is measured by connective-based F1-measure, types of discourse relations by simple ratio (or Cohen's κ).

of relation is higher for primary connectives (F1: 0.83 vs. 0.70). This is not so surprising due to the significantly bigger heterogeneity of secondary connectives (we deal with nominal, verbal, prepositional phrases etc.) in comparison with lexically frozen (i.e. grammaticalized) forms of primary connectives (the secondary allow a bigger degree of variation in terms of inflection, modification etc.). Therefore, the annotation of secondary connectives is for the human annotators more difficult.

Type of Inter-Annotator Agreement	Primary Con	Secondary Con
Existence of relation (F1)	0.83	0.70
Types of discourse relations	0.77	0.82
Types of discourse relations (Cohen's κ)	0.71	0.78

Table 1. Inter-Annotator Agreement.

On the other hand, the agreement on semantic types of discourse relations is slightly higher for secondary connectives – see simple ratios 0.77 (0.71 C. k.) vs. 0.82 (0.78 C. k.). This may be explained by the fact that most of the secondary connectives contain a transparent key word (like *condition*, *reason*, *result*, *concession*, *contrast* etc.) that refers directly to one of the individual semantic types of relations (although this relationship is not so straightforward in all cases).

In this respect, primary connectives seem to be more easily identifiable in authentic PDT texts and secondary connectives, on the other hand, signal more transparently the individual semantic types of discourse relations.

Altogether, the I-A agreement for the annotation of secondary connectives in PDT seems satisfactory (i.e. comparable with similar discourse annotation of primary connectives).

4.2 Primary vs. Secondary Connectives in Numbers

At the current stage, PDT data contain altogether 21,416 annotated discourse relations. Within this number, there are 20,255 tokens of primary connectives and 1,161 of secondary connectives – see Table 2 (the results are measured on the whole PDT data). In other words, primary connectives form 94.6 % and secondary connectives 5.4 % within the whole number of explicit discourse relations in PDT. Therefore, the terms primary and secondary connectives seem suitable also in terms of frequency – explicit discourse relations are signaled primarily by primary connectives. However, the number of secondary connectives in PDT is not insignificant and discourse annotation would be incomplete without them.

	Tokens in PDT	%
Primary Con	20,255	94.6
Secondary Con	1,161	5.4
Total	**21,416**	100

Table 2. Discourse Annotation in PDT.

The results of annotation also demonstrate that the majority of secondary connectives (924 within 1,161, i.e. 76 %) expresses discourse relations between two verbal (or clausal) arguments. The reason is that not all secondary connectives (e.g. prepositional phrases) allow nominalization of the second argument. Nominalization appears only with a set of similar structures like *výjimkou je* "the exception is", *důvodem je* "the reason is", *podmínkou je* "the condition is", *vysvětlením je* "the explanation is" etc. – see Example 4. Such secondary connectives contain the predicate already within their structure (mostly the verb *to be*) so they do not demand another finite verb in the argument and may be followed only by the nominal phrase. (The results of annotation also revealed that nominalization of the second argument even predominates in these structures – in 80 %. Thus the structure of the secondary connective has a direct influence on the syntactic realization of the second argument in these cases.)

As said above, the extended discourse annotation captures not only the secondary connectives but also the free connecting phrases (functioning as discourse indicators only in a limited set of contexts, like *kvůli jeho pozdnímu příchodu* "due to his late arrival", *kvůli tomuto nárůstu* "due to this increase", *kvůli tomuto rozhodnutí* "due to this decision" that may be mostly substituted by universal *kvůli tomu* "due to this" but not vice versa). Currently, PDT contains 1,161 tokens of secondary connectives and 151 of free connecting phrases (i.e. 88 % vs. 12 %). We may see that there is a strong tendency for multiword discourse phrases to gradually fix on one stable form and to gain a status of a universal connective.

4.3 Semantic Types of Discourse Relations

Distribution of the individual semantic relations (presented in Table 3) is for primary and secondary connectives similar, i.e. very numerous relations are relations of conjunction, reason and result and condition. The relations with the lowest (or very low) numbers are the pragmatic relations (i.e. pragmatic contrast, pragmatic reason and pragmatic condition).

On the other hand, primary and secondary connectives significantly differ in case of opposition and explication. The relation of opposition is the second most numerous relation expressed by primary connectives (with 3,171 tokens) whereas with secondary connectives, it occurred only in 13 cases. So the relation of opposition is almost exclusively expressed by primary connectives (in 99.6 %), which demonstrates that Czech does not have many multiword alternatives to signal this type of discourse relation.

On the other hand, the relation of explication is the fourth most numerous relation within secondary connectives (with 67 relations) whereas in case of primary connectives, it is in the middle (with rather low tokens within primary connectives).

However, the percentage of the individual relations clearly demonstrates that primary connectives prevail significantly in all cases (their percentage in comparison to secondary connectives is higher than 90 % in most of the relations).

Slightly higher percentage (within secondary connectives) occurs only in three types of relations: explication (22.7 %), exemplification (16.9 %) and generalization (16.7 %). However, generally, the primary connectives prevail in all the relations very clearly.

Type of Relation	Total	Primary Con	Primary Con %	Secondary Con	Secondary Con %
conjunction	7,730	7,386	95.5	344	4.1
opposition	3,184	3,171	99.6	13	0.4
reason and result	2,927	2,583	91.4	344	8.6
condition	1,451	1,351	93.1	100	6.9
concession	918	874	95.2	44	4.8
asynchronous	860	816	94.9	44	5.1
confrontation	666	632	94.9	34	5.1
specification	649	625	96.3	24	3.7
gradation	459	443	95.6	20	4.4
correction	456	439	97.1	13	2.9
purpose	419	412	98.3	7	1.7
explication	**295**	**261**	**77.3**	**67**	**22.7**
restrictive opposition	294	266	90.5	28	9.5
disj. alternative	271	228	96.3	10	3.7
synchronous	226	225	99.6	1	0.4
exemplification	**177**	**147**	**83.1**	**30**	**16.9**
generalization	**120**	**100**	**83.3**	**20**	**16.7**
equivalence	110	99	90	11	10
conj. alternative	90	88	97.8	2	2.2
pragmatic contrast	50	50	100	0	0
pragmatic reason	44	41	93.2	3	6.8
pragmatic condition	17	16	94.1	1	5.9
other	3	2	66.7	1	33.3
Total	**21,416**	**20,255**	**94.6**	**1,161**	**5.4**

Table 3. Primary vs. Secondary Connectives – Types of Discourse Relations in PDT.

4.4 New Semantic Types of Discourse Relations for Secondary Connectives

Another lesson we have learnt from the annotation of secondary connectives is that we cannot simply adopt the existing annotation principles created for the primary connectives. Secondary connectives are much more heterogeneous group than primary connectives (concerning lexical, syntactic as well as semantic aspects – see Rysová, 2012). Therefore, we can expect that it will project also to their annotation in large corpora and that the existing annotation principles will need to be modified and to react on all the differences.

As for types of discourse relations, we may expect that secondary connectives may express some new semantic relations (that are not in the classification of discourse relations formulated for primary connectives). Therefore, the human annotators were asked to mark all occurrences of secondary connectives expressing such "new" relations. Altogether, the remarks referred to three new relations: **a) entailment or deduction of results** (expressed, e.g., by secondary connectives *výsledkem je* "the result is"; *z toho vyplývá* "it follows"); **b) the relation of conclusion** (e.g. *závěrem je* "the conclusion is", *dojít k závěru* "to come to a conclusion"); **c) the relation of regard** (e.g. *v tomto ohledu* "in this respect", *v tomto směru* "in this regard"). The common feature of these relations is that they refer mostly to a larger piece of the text (e.g. to the whole previous paragraph etc.). In our opinion, these semantic relations cannot be included within any relation formulated for primary connectives and the existing classification should be extended.

4.5 Inter- and Intra-Sentential Discourse Relations

As said above, the PDT discourse annotation contains both inter- and intra-sentential relations (i.e. both *I would like to go on a trip. **But** it is raining.* and *I would like to go on a trip **but** it is raining.*). Therefore, we have analyzed whether primary and secondary connectives prefer one of these ways of expression. The ratio of inter- and intra-sentential relations expressed by primary and secondary connectives is presented in Table 4.

	Intra	%	**Inter**	%	**Total**
Primary Con	14,195	70 %	6,060	30 %	**20,255**
Secondary Con	432	37 %	729	63 %	**1,161**
Total	**14,627**	**68 %**	**6,789**	**32 %**	**21,416**

Table 4. Inter- and Intra-Sentential Relations

Table 4 demonstrates that primary connectives prefer intra-sentential discourse relations (in 70 %) while secondary connectives inter-sentential relations (in 63 %). Thus we may see that this is another crucial aspect in which primary and secondary connectives significantly differ.

We have carried out a further analysis and concentrated on the possible connection between the way of expressing discourse relations (i.e. inter- or intra-sentential) and the semantic types of given relations. We tried to examine whether this connection may give us some possible explanation why the authors prefer secondary connectives rather than primary connectives in certain contexts. We found out that in all the semantic types of relations (like reason and result, opposition etc.) prevail in both inter- and intra-sentential relations primary connectives except for two – **the inter-sentential relations of purpose and condition** prefer the expression by secondary connectives (in 86 % for purpose and 62 % for condition). This generally means that if the text contains either the inter-sentential relation of purpose or condition, there is a relatively high probability (at least in case of purpose) that they will be expressed by secondary (rather than primary) connectives.

5 Conclusion

In the paper, we have introduced the annotation of the so called secondary connectives (i.e. expressions like *the condition is*, *to conclude*, *for these reasons* etc.).

From theoretical point of view, we define discourse connectives as (mostly) universal indicators of discourse relations that may have different surface forms. According of their realization, we distinguish primary and secondary connectives. **Primary connectives** are expressions with universal status of discourse indicators that are grammaticalized (i.e. lexically frozen). They are functional words (i.e. mainly conjunctions and structuring particles) that are

not integrated into clause structure as sentence elements like *but*, *and*, *or*, *because* etc. **Secondary connectives** are mainly multiword phrases containing a lexical word or words that are not yet fully grammaticalized; therefore, these structures are much more variable (concerning modification, inflexion etc.). The secondary connectives may be sentence elements (*because of this*), sentence modifiers (*simply speaking*) or they form a separate sentence (*The reason is simple*.).

In the paper, we demonstrated how it is possible to include secondary connectives into corpus annotations. The overall inter-annotator agreement on existence of a discourse relation is 0.70 (F1) and on the type of a discourse relation 0.82 (0.78 C. k.), which is very similar to primary connectives in PDT.

Altogether, PDT contains 1,161 tokens of secondary connectives, which is 5.4 % within all explicit discourse connectives in PDT (thus the attribute secondary seems suitable for them also in terms of frequency).

We have compared primary and secondary connectives also in terms of semantic types of discourse relations they express. The distribution of the individual semantic relations is very similar for both primary and secondary connectives (with some exceptions like the relation of opposition occurring very predominantly with primary connectives). However, the annotation has taught us that the classification of relations formulated for primary connectives cannot be simply adopted for secondary connectives – during the annotation, we have observed three "new" semantic types (that were not included into the classification for primary connectives): a) **entailment or deduction of results** (e.g. *it follows*); b) **the relation of conclusion** (e.g. *the conclusion is*); c) **the relation of regard** (e.g. *in this respect*). These three types of relation refer mostly to larger pieces of text like a whole paragraph.

The results of annotation also demonstrate that primary and secondary connectives differ in terms of inter- and intra-sentential relations. Whereas primary connectives prefer the intra-sentential relations (in 70 %), secondary connectives mostly the inter-sentential relations (in 63 %). So primary and secondary connectives do not differ only from syntactic, lexical and semantic point of view, but also in the way how they structure the text.

At the current stage, the Prague Dependency Treebank contains the most detailed annotation of secondary connectives (as far as we know, done on the largest data) that could be adopted also for other languages in other corpora focusing mostly on the annotation of primary connectives. In the paper, we tried to demonstrate that discourse annotation including secondary connectives is more complete and that similar analysis may lead to better understanding of discourse.

Acknowledgement

This paper was supported by the project "Discourse Connectives in Czech" (n. 36213) solved at the Faculty of Arts at the Charles University in Prague from the resources of the Charles University Grant Agency in 2013–2015.

The authors acknowledge support from the Czech Science Foundation (project n. P406/12/0658) and from the Ministry of Education, Youth and Sports of the Czech Republic (project n. LH14011 and LM2010013).

This work has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).

The authors gratefully thank to Jiří Mírovský from the Charles University in Prague for his kind measuring the figures on the PDT data for this paper.

References

Karin Aijmer. 2002. *English Discourse Particles. Evidence from a corpus*. Studies in Corpus Linguistics 10. Amsterdam/Philadelphia: John Benjamins, ISBN 90-272-2280-0.

Nicholas Asher. 1993. *Reference to Abstract Objects in Discourse*. Dordrecht: Kluwer Academic Publishers, ISBN 0-7923-2242-8.

Eduard Bejček et al. 2012. Prague Dependency Treebank 2.5 – a revisited version of PDT 2.0. In: *Proceedings of Coling 2012*, Bombay, India, pp. 231–246.

Teuen A. van Dijk. 1979. Pragmatic Connectives. In: *Journal of Pragmatics 3*. North-Holland Publishing Company, pp. 447–456.

Kerstin Fischer, eds. 2006. *Approaches to Discourse Particles*. Studies in Pragmatics 1. Amsterdam: Elsevier, ISBN-10: 0080447376, ISBN-13: 978-0080447377.

Bruce Fraser. 1999. What are discourse markers? *Journal of Pragmatics 31 (7)*. Elsevier, pp. 931–952.

Maj-Britt Mosegaard Hansen. 1998. *The Function of Discourse Particles: A study with special reference to spoken standard French*. Amsterdam: John Benjamins. ISBN 9789027250667.

Douglas Harper. 2001. Online Etymology Dictionary. <http://www.etymonline.com>.

Lucie Poláková et al. 2013. Introducing the Prague Discourse Treebank 1.0. In: *Proceedings of the 6th International Joint Conference on Natural Language Processing*, Copyright © Asian Federation of Natural Language Processing, Nagoya, Japan, ISBN 978-4-9907348-0-0, pp. 91–99.

Lucie Poláková et al. 2012. *Prague Discourse Treebank 1.0*. Data/software, ÚFAL MFF UK, Prague, Czech Republic.

Rashmi Prasad et al. 2008. The Penn Discourse Treebank 2.0. In: *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, pp. 2961–2968.

Rashmi Prasad, Aravind Joshi, Bonnie Weber. 2010. Realization of Discourse Relations by Other Means: Alternative Lexicalizations. In: *Proceedings of Coling 2010*, Tsinghua University Press, Beijing, China, pp. 1023–1031.

Rashmi Prasad, Bonnie Webber and Aravind Joshi. 2014. Reflections on the Penn Discourse TreeBank, Comparable Corpora and Complementary Annotation. *Computational Linguistics 40 (4)*, pp. 921–950.

Magdaléna Rysová. 2012. Alternative Lexicalizations of Discourse Connectives in Czech. In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), European Language Resources Association, Istanbul, Turkey, ISBN 978-2-9517408-7-7, pp. 2800–2807.

Magdaléna Rysová, Jiří Mírovský. 2014. Use of Coreference in Automatic Searching for Multiword Discourse Markers in the Prague Dependency Treebank. In: Proceedings of The 8th Linguistic Annotation Workshop (LAW-VIII), Copyright © Dublin City University (DCU), Dublin, Ireland, ISBN 978-1-941643-29-7, pp. 11–19.

Magdaléna Rysová, Kateřina Rysová. 2014. The Centre and Periphery of Discourse Connectives. In: Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing, Copyright © Department of Linguistics, Faculty of Arts, Chulalongkorn University, Bangkok, Thailand, ISBN 978-616-551-887-1, pp. 452-459.

Deborah Schiffrin. 1987. *Discourse markers*. Cambridge: Cambridge University Press, ISBN 9780521357180.

Manfred Stede, Arne Neumann. 2014. Potsdam Commentary Corpus 2.0: Annotation for discourse research. In: N. Calzolari et al. (eds.), *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavík: European Language Resources Association (ELRA), pp. 925–929.