Verbs of Saying with a Textual Connecting Function in the Prague Discourse Treebank

Magdaléna Rysová

Charles University in Prague, Faculty of Arts

nám. Jana Palacha 2, 116 38 Praha

magdalena.rysova@post.cz

Abstract

The paper tries to contribute to the general discussion on discourse connectives, concretely to the question whether it is meaningful to distinguish two separate groups of connectives – i.e. "classical" connectives limited to few predefined classes like conjunctions or adverbs (e.g. *but*) vs. alternative lexicalizations of connectives (i.e. unrestricted expressions and phrases like *the reason is*, *he added*, *the condition was* etc.). In this respect, the paper focuses on one group of these broader connectives in Czech – the selected verbs of saying *doplnit/doplňovat* (*to complement*), *upřesnit/upřesňovat* (*to specify*), *dodat/dodávat* (*to add*), *pokračovat* (*to continue*) – and analyses their occurrence and function in texts from the Prague Discourse Treebank. The paper demonstrates that these verbs of saying have a special place within the other connectives, as they contain two items – e.g. *he added* means *and he said* so the verb *to add* contains an information about the relation to the previous context (*and*) plus the verb of saying (*to say*). This information led us to a more general observation, i.e. discourse connectives in broader sense do not necessarily connect two pieces of a text but some of them carry the second argument right in their semantics, which "classical" connectives can never do.

Keywords: connectives, discourse relations, verbs of saying

1. Motivation

The paper is based on the annotation of discourse (i.e. textual) relations in the Prague Discourse Treebank (PiDT). It examines the possibilities of how discourse relations may be realized, i.e. which language means have an ability to signal that a certain part of a text is related to another.

Discourse relations in Czech are already annotated in the Prague Discourse Treebank (cf. Poláková et al., 2012 – it will soon enrich the new version of the Prague Dependency Treebank /Bejček et al., 2012/). However, this version of the corpus captures only such relations that are expressed by explicit discourse connectives – understood (in PDiT approach) as expressions of certain pre-defined classes – mainly conjunctions, adverbs and particles (like *and*, *or*, *but*, *then*, *however*, *only* etc.). Nevertheless, the annotation revealed that some of these relations are realized also by other means – i.e. expressions with a connecting function at the layer of discourse that are both syntactically and lexically unrestricted.

2. Aim of the Paper

The aim of the paper is to analyse one specific group of such alternative expressions with a connecting function, namely verbs of saying that were identified during the annotation of the textual relations in the Prague Discourse Treebank – i.e. Czech verbs *doplnit/doplňovat* [1] (*to complement*), *upřesnit/upřesňovat* (*to specify*), *dodat/dodávat* (*to add*), *pokračovat* (*to continue*). These verbs have been captured by annotators during the preliminary manual annotation, so the analysis focuses on them. The aim of the paper is not to describe all Czech verbs of saying.

On the basis of the data analysis, the paper tries to contribute to a more general issue, i.e. to specify the difference between "classical" connectives (like *because*, *but*) and verbs of saying with a connecting function. Therefore, the present paper tries to solve this theoretical question with the aim to use the results for practical annotations of enriched textual relations of the Prague Discourse Treebank.

At the same time, these general observations may be helpful also for other treebanks like Penn Discourse Treebank for English or Potsdam Commentary Corpus for German containing textual annotations, in terms of how it is possible to capture textual relations expressed by verbs of saying.

3. Broader Possibilities of Expressing Textual Relations – Alternative Lexicalizations of Discourse Connectives

The analysis is carried out on the data of the Prague Dependency Treebank (Bejček et al., 2012) – a large corpus that contains annotation of more levels at once (morphological, syntactic and underlying syntactico-semantic called tectogrammatical). Moreover, tectogrammatical level was (as in the only corpus of Czech) enriched also by annotation of textual relations – published independently as the Prague Discourse Treebank (Poláková et al., 2012). The first phase of textual annotation captured only relations expressed by "classical" connectives like *therefore*. Nowadays this annotation is being enriched by annotation of broader or alternative possibilities of expressing textual relations.

Existence of these broader possibilities of signalling textual relations on a large corpus data was described first for English (however, the study does not deal with verbs of saying) on the data of the Pennsylvanian Penn Discourse Treebank (cf. Prasad et al., 2010) – the authors of the study called these expressions alternative

---

[1] The two forms represent different types of verbal aspect – perfective and imperfective.

lexicalizations of discourse connectives (shortly AltLex's) – e.g. *that means*; *one reason is* etc.

The first probe of these alternative ways of expressing textual relations was done for Czech on the data of the Prague Discourse Treebank (cf. Rysová, 2012). However, this analysis was carried out on a small sample of data (i.e. 306 AltLex's that have been captured during the preliminary manual annotation) and the aim was to point at the existence of such expressions also in Czech.

At the same time, it is very complicated to clearly define the wide category of AltLex's and to delimit their boundaries. They oscillate between one-word (e.g. *přeloženo* – in English *translated*) and sentential expressions (e.g. *Důvod je jednoduchý* – in English *The reason is easy.*). They may be both grammatically and lexically restricted (e.g. *jednoduše řečeno – simply saying* – it means that this expression functions as AltLex only in this form; the verb *říct* /*to say*/ and the adverb *jednoduše* /*simply*/ do not have a connecting function on their own) or unrestricted (e.g. the verb *následovat – to follow* function as AltLex in its whole paradigm). Some of them have a noun as the core of their lexical meaning – e.g. *důvod* (*reason*) occurring in several combinations like *důvodem je* (*the reason is*) or *jako důvod uvedla* (*she gives the reasons*). Some of them are verbal – cf. *dodat* (*to add*), *specifikovat* (*to specify*), *pokračovat* (*to follow*) etc.

It is obvious that the category of AltLex's is very broad and complex and that it is necessary to study them in more detail through the individual groups. The present paper tries to examine one such group – verbs of saying introducing (in)direct speech that were identified during the preliminary manual annotation in the Prague Discourse Treebank (cf. Rysová, 2012).

4. Verbal Alternatives of Discourse Connectives in the Prague Discourse Treebank

One of the most numerous groups (according to the new and enriched annotation, it is approximately 2,200 out of 5,000 tokens AltLex's in PDiT) [2] are verbal AltLex's, i.e. those expressions with connecting function having the verb as their basis. At the same time, it is the verb signalling the type of textual relation – e.g. the verb *odůvodnit* (*to give reasons*) signals a relation of reason and result, *specifikovat* (*to specify*) most often a relation of specification etc.

The paper presents the analysis of one subgroup of verbal Czech AltLex's – four verbs of saying introducing (in)direct speech identified during the preliminary textual annotation of the Prague Discourse Treebank [3]. However, the aim of the paper is not to describe all of these verbs but to show some general tendencies on the selected representatives.

4.1 Verbs of Saying Introducing (In)Direct Speech

Within 49,431 of sentences in the preliminary manual annotation of AltLex's in PDiT, the annotators identified four verbs of saying introducing (in)direct speech that have a connecting function within a text. These verbs are *doplnit/doplňovat* (*to complement*), *upřesnit/upřesňovat* (*to specify*), *dodat/dodávat* (*to add*), *pokračovat* (*to continue*) – cf. Table 1 with the number of tokens in the Prague Discourse Treebank.

Table 1 demonstrates that there are 558 instances of the selected verbs in PDiT, all manually annotated.

All of these verbs are polysemantic so they introduce the (in)direct speech only in some instances (in PDiT, it is approximately half of all – 270).

In most of the other meanings, the selected verbs do not function as AltLex's at all, it means that they do not have a connecting function – e.g. the verb *dodat/dodávat* (*to add*) may also mean '*to supply, to deliver*' – like *dodat pivo do restaurace* (*to deliver beer to the restaurant*). The similar instances clearly do not function as indicators of discourse relations.

Only the verb *pokračovat* (*to continue*) has also another AltLex meaning than introducing (in)direct speech. In one of its meanings, it introduces the discourse relation of precedence and succession like *v říjnu začal hospodařit a pokračuje dodnes* (*he began to farm in October and continues up today*). However, the other meanings of the selected verbs (whether AltLex or not) are not the main topic of this paper and we leave them aside. In the rest of the paper, we will focus only on the selected verbs of saying in the function of AltLex's introducing the (in)direct speech – like Example (1) from the Prague Discourse Treebank:

(1) *S kolegy jsem se seznámil až po prvním dějství, řekl Peter Dvorský.*
*Potom jsem měl plný kalendář, **dodal**.*

(*I got to know my colleagues after the first act, said Peter Dvorský.*
*Then I had a full schedule, he **added**.*)

In PDiT, there are altogether 270 instances of selected verbs introducing the (in)direct speech. It is interesting that most of them (234) are represented by the verb *dodat/dodávat* (*add*). It seems that the function as a verb of saying is dominant for this verb, as it introduces (in)direct speech in 77 % of its instances in PDiT. The instances of the other verbs introducing (in)direct speech are not so numerous and at the same time the percentage of this function is also not so high in their case – cf. *upřesnit/upřesňovat* (*specify*) 14 instances as a verb of saying, which is 56 % of all its instances; *doplnit/doplňovat* (*complement*) 9 instances, 19 %; *pokračovat* (*continue*) 13 instances, 7 %. From this observation, it is obvious that some of these verbs

---

[2] These figures are not final, as the new enriched annotation of these expressions has not been finished yet. They are, therefore, an approximation done on the basis of the so far annotated part of the corpus.

[3] It is possible that there are more such verbs of saying in PDiT that were not captured by the first group of annotators; the new and more detailed annotation is now in progress.

function as verbs of saying dominantly (*dodat/dodávat /add/*), some only marginally (cf. the verb *pokračovat /to continue/* occurring in PDiT in relatively numerous instances – 179, within which only in 7 % as a verb of saying).

Lemma	Introducing (In)Direct Speech	Other Meanings	Total Tokens in PDiT
dodat/dodávat (*add*)	234	72	306
doplnit/doplňovat (*complement*)	9	39	48
pokračovat (*continue*)	13	166	179
upřesnit/upřesňovat (*specify*)	14	11	25
TOTAL	**270**	**288**	**558**

Table 1: Verbs of Saying Introducing (In)Direct Speech in the Prague Discourse Treebank

4.2 Annotation of Verbs of Saying Introducing (In)Direct Speech in the Prague Discourse Treebank

The chosen verbs of saying have been manually annotated in the Prague Discourse Treebank and they have been labelled as AltLex's (i.e. that they have a connecting function within a text) – they carry certain meaning that presupposes a presence of some other components in the text (which is a general feature of all connective means – cf. Halliday and Hasan, 1976). In other words, these verbs usually do not appear in the first sentence of a text, as they imply a presence of another verb of saying in the previous part of the text – e.g. the sentence *John added* implies that John (or someone else) said also something before. This aspect connects these Czech verbs of saying with "classical" connectives, which means that they may be also considered indicators of textual relations, as they are involved into the constitution of a text – see Example (1).

In this example, there is a relation of conjunction[4] between the verbs *říct* (*to say*) and *dodat* (*to add*). This is obvious from the fact that *he added* means in fact '*and he said*'. In other words, the verb *to add* contains an information about the relation to the previous context (*and*) plus the pure verb of saying (*to say*). From this reason, *to add* (similarly as "classical connectives") usually do not stand in the first sentence in the text, as it implies that someone said something before.

At the same time, there is also another textual relation in Example (1) than conjunction between verbs of saying – there is also a relation of precedence and succession within the indirect speech, i.e. between *I got to know my colleagues after the first act* and *I had a full schedule* expressed by the connective *then* (translated from the original).

Therefore, there are two layers of textual relations –

one within the introducing verbs of saying and the second within the contents of the (in)direct speech – see another example from PDiT:

(2) *Zdůraznil, že výsledek hlasování zavazuje celou Francii, neboť, jak dodal, z vítězství nevyšel žádný Francouz ani jako vítěz, ani jako poražený, ať hlasoval jakkoliv.*

(*He stressed that the outcome of the vote commits the whole France because, as he added, each Frenchman is neither a winner nor loser, whoever he voted.*)

In the example, there are two textual relations. The first is a relation of conjunction between the verbs of saying signalled by the verb *dodat* (*to add*) (in other words, it is *he stressed and he said*), the second is reason and result within the indirect speech (*the outcome of the vote commits...* and *each Frenchman is*) expressed by the connective *neboť* (*because*).

In this respect, we have annotated all of the selected verbs of saying – see Table 2.

Table 2 demonstrates the annotation of selected verbs of saying in the Prague Discourse Treebank.

The column "AltLex" contains examples where the discourse relation is expressed by an AltLex verb of saying – see Examples (1) and (2) with discourse relation of conjunction signalled by AltLex phrases *dodal* (*he added*) and *jak dodal* (*as he added*).

The Prague Discourse Treebank captures discourse relation between two parts of a text, in PDiT terminology between two verbal arguments. The column "One argument" contains instances where the first verbal argument of a discourse relation is missing (the author uses a connective expression for seeming evocation of better continuity of a text or reacts to some nonverbal situation) or the first argument is not expressed by a verb – see Example (3):

---

[4] The types of textual relations (e.g. conjunction, reason and result, opposition etc.) are assigned to the individual examples in agreement with the manual for annotation of textual relations in the Prague Discourse Treebank (cf. Mladová et al., 2012).

(3) _Podle slov_ ředitele Pomezného by se měla podpora vývozu především na vládní úrovni v budoucnu zpřehlednit.

Proexportní politika však vždy bude kombinací vládních a nevládních iniciativ, **dodal** _Pomezný._

(_According to the words_ of the director   Pomezný, the export support should be clarified in the future especially at the government level.

However, the pro-export policy is always a combination of governmental and non-governmental initiatives, **added** _Pomezný._)

In Example (3), the first indirect speech (_the export support should be..._) is not introduced by a verb of saying like _ředitel řekl_ (_the director said_) but by a prepositional phrase _podle slov_ (_according to the words_). We believe that the discourse relation is between the two phrases of saying introducing (in)direct speech (see section 5). However, in this case, the (in)direct speech is not introduced by a verb but by a prepositional phrase _podle slov_ (_according to the words_). Therefore, PDiT does not annotate here (in the current stage) any relation (as it captures only discourse relations between verbal arguments) and only provides these examples with the note "one argument" (from this reason, these cases are also not included into the section 4.3 introducing the individual types of discourse relations).

Verbs Introducing (In)Direct Speech	AltLex	One argument	TOTAL
dodat/dodávat (_add_)	189	45	**234**
doplnit/doplňovat (to complement)	5	4	**9**
pokračovat (_continue_)	13	0	**13**
upřesnit/upřesňovat (_specify_)	14	0	**14**
TOTAL	**221**	**49**	**270**

Table 2: Annotation of Chosen Verbs of Saying in the Prague Discourse Treebank

4.3 Types of Discourse Relations Expressed by Verbs of Saying Introducing (In)Direct Speech in the Prague Discourse Treebank

We have also analysed the selected verbs of saying in terms of the type of discourse relations they express.

It may be supposed that the meaning of the verb is relatively transparent, i.e. that the verb _specifikovat_ (_to specify_) signals mostly the relation of specification, the verb _dodat/dodávat_ (_to add_) the relation of conjunction etc. The final results of the manual annotation are demonstrated in Table 3.

Types of Discourse Relation	Verbs Introducing (In)Direct Speech				TOTAL
	dodat/dodávat (_add_)	_doplnit/doplňovat_ (_complement_)	_pokračovat_ (_continue_)	_upřesnit/upřesňovat_ (_specify_)	
Conjunction	185	5	13		**203**
Specification				10	**10**
Equivalence				1	**1**
Explication				3	**3**
Opposition	3				**3**
Concession	1				**1**
TOTAL	**189**	**5**	**13**	**14**	**221**

Table 3: Types of Discourse Relations Expressed by Verbs of Saying

Table 3 demonstrates which types of textual relations are expressed by the selected verbs. The verbs *dodat/dodávat* (*to add*), *doplnit/doplňovat* (*to complement*) and *pokračovat* (*to continue*) express in most cases the relation of conjunction – see examples (1) and (2), the verb *upřesnit/upřesňovat* (*to specify*) the relation of specification. Only four instances of *dodat/dodávat* (*to add*) signal opposition or concession. However, all of them are instances of a special larger phrase *nutno/dlužno dodat* (*it is necessary to add*) that differ from the other instances of *dodat/dodávat* (*to add*), as this phrase does not connect two verbs of saying – see example (4):

(4) *Novináři jsou hlídací psi společnosti, nezávislý kontrolní orgán uvnitř státu, prostě sedmá velmoc. Taková je všeobecně sdílená představa o poslání novinářů.*
Dlužno dodat, že nikdo se na vytvoření tohoto obrazu nepodílel právě tak jako sami novináři.

(*Journalists are the watchdogs of the society, the independent supervisory authority within the state, just the seventh power.*
This is a widely shared vision of the mission of journalists.
It is necessary to add that no one has been involved in the creation of this image more than the

In Example (4), the phrase *dlužno dodat* (*it is necessary to add*) functions as an indicator of the opposition between *this is a widely shared vision...* and *no one has been involved in the creation...* The phrase *dlužno dodat* (*it is necessary to add*) is replaceable by some of the connectives of the opposition like *ovšem* (*however*) or *ale* (*but*). In some cases in PDiT, this phrase and some of these connectives occur even together strengthening the relation of opposition – *nutno ovšem dodat* (*it is, however, necessary to add*).

4.4 Interplay of Syntactic and Discourse Level – Connective Raising

During the analysis of the chosen verbs of saying, we observed a special structure from the syntactic and semantic point of view. As said in the section 4.2, most instances of verbs of saying contain two levels of discourse relations – the one between the verbs of saying themselves (e.g. *he said* and *he added*) and the second within the contents of the (in)direct speech expressed by some other means, e.g. connectives like *potom* (*then*) (see Example 1) or *neboť* (*because*) (see Example 2). In most cases, these connectives are embedded in the subordinate clause, i.e. they belong to the content of the (in)direct speech – see an Example (1) – *I got to know my colleagues after the first act, said Peter Dvorský. Then I had a full schedule, he added.* The connective *then* indicating the relation of precedence and succession belongs to the subordinate clause *I had a full schedule*. In these cases, the syntax goes hand in hand with semantics.

However, there is another structure where syntax

and semantics go against each other – see Example (5):

(5) *Po jmenování uvedli, že pokud to bude nutné, pozastaví své členství v mateřských stranách.*
Jak však dodali, nezávislost není zaručena vystoupením ze strany.

(*They said after the appointment that, if necessary, they will suspend their membership in the parent parties.*
But they added that the independence is not guaranteed by secession from the party.)

Again, in Example (5), there are two levels of discourse relations. The first relation of conjunction is between the two verbs of saying *uvedli* (*they said*) and *dodali* (*they added*); the second relation of opposition is between the contents of the indirect speeches, i.e. between *they will suspend their membership...* and *the independence is not guaranteed*. This relation of opposition is expressed by the connective *však* (*but*). However, the connective *však* (*but*) is not embedded into the subordinate clause but is raised to the level of the main clause. So syntactically, the connective *však* (*but*) is a part of the main clause (*jak dodali /they added/*) but semantically, it belongs to the lower level of the subordinate clause.

It is interesting, that this phenomenon occurs (at least in the Prague Discourse Treebank) only with discourse relation of opposition and connectives *ovšem/však* (*however*) and *ale* (*but*).

5 Difference between "Classical" Connectives and AltLex's – General Reflection

The above analysis of verbs of saying has led us to further thinking about the general difference between "classical" connectives (like *therefore*, *but*, *and*) and their alternative lexicalizations (i.e. AltLex's). As said above, connectives are (in the PDiT approach but also by some other authors like Halliday and Hasan, 1976; Martin, 1992; Knott, 1996) understood as expressions from some pre-defined classes (mainly adverbs, connectives, particles), i.e. they are fixed and the list of connectives is limited.

One crucial point about connectives is that they express a textual relation between two units or parts of the text (PDiT uses the term arguments), i.e. the connectives stand outside or above them. Textual relation may be then described as a relation between argument one (ARG1) and argument two (ARG2) expressed by a connective (CONN) – see Example (6) from PDiT:

(6) *Přestup do Evropy je pro každého hráče z Jižní Ameriky velkým krokem do neznáma.* ARG1
Proto **CONN** *si musí najít klub, kde mají pro jeho aklimatizační problémy pochopení.* ARG2

(*Transfer to Europe is a big step into the unknown for each player from South America.* ARG1
Therefore **CONN**, *he must find a club tolerant of the*

acclimatization problems. ARG2)

In Example (6), there is a relation of reason and result between argument one (*transfer to Europe is...*) and argument two (*he must find...*) signalled by the connective *therefore*.

As discussed in section 3, AltLex's are much more diverse – lexically, syntactically and semantically (cf. Rysová, 2012). Moreover, we want to demonstrate here that they are diverse also in terms of the division of participants of textual relation (i.e. with regard to the concept of ARG1_CONN_ARG2).

Within Czech AltLex's, there are expressions that are fully replaceable by "classical" connectives, so their basic function is to connect two textual arguments in a certain type of textual relation – cf. expressions like *to je důvod, proč* (*that is the reason why*); *because of this*; *z tohoto důvodu* (*from this reason*) that are replaceable by the connective *proto* (*therefore*) and the meaning remains the same – see Example (6). From this point of view, they also stand between two textual arguments.

However, the above analysis of verbs of saying revealed a specific group of Czech AltLex's that differ from the others exactly in this way. In other words, there are some verbal AltLex's that have the function of a connective and the argument two at once. In this way, we may say that the verb *dodat* (*to add*) combines both the connective *a* (*and*) and the argument two *říct* (*to say*), i.e. *he added = and he said = and + to say*. Therefore, there is a difference between "classical" connectives – see Example (7) – and a specific verbal group of AltLex's – see Example (8):

(7) *He said* ARG1 *something*. **And** **CONN** *he said* ARG2 *something.*

(8) *He said* ARG1 *something*. *He added* **CONN+ARG2** *something*.

We believe that this observation may contribute to the specification of the difference between connectives and one group of their alternative lexicalizations in general (i.e. what connectives cannot do and some AltLex's can do), as well as to the description of such broad and diverse AltLex category whose boundaries is hard to define.

6   Conclusion

In the present paper, we examine one specific group of Czech alternative lexicalizations of discourse connectives – the four verbs of saying captured by the preliminary textual annotation in the Prague Discourse Treebank. Based on our analysis, we try to draw some general observations about the specific ability of this group of Czech AltLex's that distinguishes them from "classical" connectives. In particular, connectives in a general case stand between two textual arguments whereas the chosen verbs of saying include the "connective" and the second argument at once (e.g. *to add* = '*and + to say*'), which

"classical" connectives and other AltLex's cannot do. Therefore, it is important to distinguish these verbs of saying as a special category within other connective means (whether connectives like *therefore* or AltLex's like *the reason is*). This general observation may help to annotate these expressions properly in the textual annotation in treebanks like the Prague Discourse Treebank or Penn Discourse Treebank.

8   References

Bejček, E. et al. (2012). Prague Dependency Treebank 2.5 – a revisited version of PDT 2.0. In: *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012)*. Mumbai, India.

Halliday, M. A. K.; Hasan, R. (1976). *Cohesion in English*. London: Longman.

Knott, A. (1996). *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. Edinburgh: University of Edinburgh, Ph.D. thesis.

Martin, J. R. (1992). *English text: System and structure*. Amsterdam: Benjamins.

Mladová, L.; Zikánová, Š.; Bedřichová, Z.; Mírovský, J.; Jínová, P.; Zdeňková, J.; Rysová, M.; Hajičová, E. (2012). *Manual of Annotation of Discourse Relations in Prague Dependency Treebank*. Prague: ÚFAL MFF UK.

Poláková, L. et al. (2012). *Prague Discourse Treebank 1.0* [CD-ROM]. Prague: ÚFAL MFF UK.

Prasad, R. et al. (2010). Realization of Discourse Relations by Other Means: Alternative Lexicalizations. In: *Coling 2010: Posters*, pp. 1023–1031.

Rysová, M. (2012). Alternative Lexicaliztions of Discourse Connectives in Czech. In Calzolari, Nicoletta et al. (eds.): *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turecko: European Language Resources Association (ELRA), pp. 2800–2807.