Alternative Lexicalizations of Discourse Connectives in Czech

Magdaléna Rysová

Charles University in Prague

Faculty of Mathematics and Physics

Institute of Formal and Applied Linguistics

Malostranské náměstí 25, 118 00 Praha 1

magdalena.rysova@post.cz

Abstract

The paper concentrates on which language means may be included into the annotation of discourse relations in the Prague Dependency Treebank (PDT) and tries to examine the so called alternative lexicalizations of discourse markers (AltLex's) in Czech. The analysis proceeds from the annotated data of PDT and tries to draw a comparison between the Czech AltLex's from PDT and English AltLex's from PDTB (the Penn Discourse Treebank). The paper presents a lexico-syntactic and semantic characterization of the Czech AltLex's and comments on the current stage of their annotation in PDT. In the current version, PDT contains 306 expressions (within the total 43,955 of sentences) that were labeled by annotators as being an AltLex. However, as the analysis demonstrates, this number is not final. We suppose that it will increase after the further elaboration, as AltLex's are not restricted to a limited set of syntactic classes and some of them exhibit a great degree of variation.

Key words: alternative lexicalization of discourse markers (AltLex); discourse connectives; discourse relations

1.   Introduction

The paper results from the annotation of textual (discourse) relations in the Prague Dependency Treebank (PDT). In particular, we intend to demonstrate how the annotation of the relations going "beyond the sentence" may be used for further theoretical as well as empirical research.

One of the ways how to annotate discourse relations is based on the identification of the so called discourse relation markers (DRMs) or connectives. However, there is not any clear and uniform definition of this category and, therefore, there is rather a general intuitive understanding what an DRM actually is. Some authors (e.g., Halliday and Hasan, 1976; Martin, 1992; Knott, 1996) define DRMs as fixed expressions out of a few well-defined syntactic classes (conjunctions, adverbs, prepositional phrases), while Prasad et al. (2010) oppose that on the basis of this definition, "literature presents lists of DRMs, which researchers try to make as complete as possible for their chosen language" and they argue that DRMs are not a closed but an open-ended class.

In the course of the annotation of the Penn Discourse Treebank (PDTB), Prasad et al. (2010) have found a wide range of additional expressions called alternative lexicalizations (AltLex) that have the same function as "classic" connectives. These expressions also signal some relation between two arguments but their lexico-syntactic nature is different from DRMs. Examples of AltLex's could be *a major reason is...*; *that may be because...*; *a consequence of their departure could be...* It seems that the annotation of AltLex's enlarges the class of expressions denoting the discourse relation into endless dimensions.

The aim of the present paper is to examine - on the basis of Prasad et al.'s research - the possible class of alternative lexicalizations in Czech. These expressions are already annotated under the discourse relations in the Prague Dependency Treebank (PDT), but their annotation is only in its beginnings. The annotators are encouraged to mark any expression that signals some relation to the previous argument (and is not a "classic" connective) with the comment "AltLex". In our paper, we present the results of our search for these expressions for the upcoming version of PDT (the search was done on the data in the current version PDT 2.0) and we discuss whether Prasad et al.'s characterization and description of English AltLex's is suitable also for Czech – cf. the following example:

(1)   *The Brazilian football player attacked his opponent in today's match. <u>This is the reason why</u> he will not play in the next three matches.*

Hráč brazilského týmu napadl v dnešním utkání svého protihráče. <u>To je důvod, proč</u> nebude hrát příští tři zápasy.

The AltLex is here *this is the reason why* (*to je důvod, proč*) and, in fact, it is replaceable by the connective *therefore*. Cf.:

(2)   *The Brazilian football player attacked his opponent in today's match. Therefore, he   will not play in the next three matches.*

It is obvious that both utterances are in the relation of "reason – result". This example thus clearly demonstrates that the relation "reason – result" may be expressed by both ways, either by a "classic" connective (*therefore*) or by AltLex (*this is the reason why*) in both languages.

2.   AltLex's in PDT

Altogether, we have found 306 tokens in the total 43,955 of sentences in PDT that were provided with the annotators' comment AltLex. However, this number is rather approximate, as some of the tokens were

misinterpreted (i.e. they were not AltLex's because they did not signal any discourse relation). Therefore, the number of tokens was finally reduced to 261.

On the other hand, there are definitely other AltLex's in PDT that were not labeled at all. As an example, we looked up one type of AltLex's (the expressions containing the word *speaking – řečeno*) and we found out how they are annotated. Among them, there were such expressions like *shortly speaking* (*stručně řečeno*) or *simply speaking* (*jednoduše řečeno*). Altogether, the upcoming version of PDT contains 53 of these expressions out of which 23 are used as discourse markers and, therefore, they should be captured in the annotation. However, their annotation varies. 3 of these expressions were labeled as AltLex's, 7 as "classic" connectives and the rest (13) remained without annotation – see Table 1:

| Expression | Total number | Use as a Discourse Marker | Annotated | | Unannotated |
			as Connective	as AltLex	
{*simply, shortly, generally...*} *speaking*	53	23	7	3	13

Table 1: Annotated and Unannotated examples of AltLex's

It means that the current annotation of the Czech AltLex's is rather inconsistent. The reason is that the annotation of these expressions is in the first phase and the research is in progress. A more elaborated annotation of AltLex's is intended for one of the future versions of PDT.

3. Classification of AltLex's

Among the 261 tokens, there were 94 types of AltLex's. We have carried out their lexico-syntactic and semantic characterization and compared it with the characterictics of the English AltLex's from PDTB 2.0.

3.1 Lexico-syntactic characterization

Prasad et al. describe the English AltLex's on the basis of syntactic and lexical flexibility. The first parameter is whether the expression belongs to one of the syntactic classes admitted as explicit connectives in the PDTB (i.e. subordinating conjunctions, coordinating conjunctions, prepositional phrases and adverbs). The second parameter examines the AltLex's in terms of lexical "stability", i.e. whether the expression is frozen or open-ended. On the basis of these criteria, Prasad et al. suggest a further subdivision of English AltLex's into three groups: 1) syntactically admitted, lexically frozen (*for one thing*), 2) syntactically free, lexically frozen (*that is why*), 3) syntactically and lexically free (*that compares with*). The authors also present the basic English AltLex patterns – they argue that AltLex's from the group 3 are modifiable and may have their core plus obligatory and optional elements like noun phrases (NX), prepositional phrases (PPX), verb phrases (VX) or adjectival phrases (JJX). An example of such complex AltLex and its pattern would be: *... attributed the increase to... = attributed <NX> to*. Therefore, some AltLex's found in PDTB (e.g. *a major reason is*) may be realized also by other variants (e.g. *the reason is*, *a possible reason for the increase is* etc.)

We have tried a similar classification for Czech AltLex's from PDT. Syntactic classes admitted for "classic" connectives are the following: coordinating conjunctions, subordinating conjunctions, particle expressions (including rhematizers), adverbs, some uses of pronouns, fixed multiple-word expressions with linking function, elements formed by letters or numbers expressing enumeration (Mladová et. al, 2011). We may see that there are more admitted syntactic classes for connectives in PDT than in PDTB. Therefore, the delimitation of connectives and AltLex's slightly differs in these two approaches. Whereas PDTB has a more limited space for connectives and a broader for AltLex's, the opposite is true for PDT. Obviously, this fact became evident also in the percentage of the AltLex expressions from the admitted syntactic classes: whereas PDTB contains 14.7 % of them out of 624 AltLex tokens (Prasad et. al, 2010), their number in PDT is higher – 26 % out of 261 tokens – see Table 2[1]:

---

[1] "Syntactically admitted" means that the expression belongs to one of the syntactic classes admitted for Czech connectives in PDT; other syntactic classes are called "syntactically free". This terminology is adopted from the study on English AltLex's in PDTB (Prasat et. al, 2010).

	Syntactically admitted	%	Syntactically free	%	Total
Number of AltLex types	37	39	57	61	94
Number of AltLex tokens	68	26	193	74	261

Table 2: AltLex's from syntactically admitted and syntactically free classes

If the expression labelled as an AltLex belongs to one of the syntactically admitted classes, it means that it should be re-annotated as a connective. This proves that discourse connectives (at least in the PDT approach) should not be understood as a close category but rather as an open-ended class of expressions. However, these expressions are treated as AltLex's in the current stage of PDT annotation and, therefore, they are called AltLex's also in the rest of this paper.

The task for the future work also is whether the boundary between the Czech connectives and AltLex's is placed suitably. The main disputable class seems to be fixed multiple-word expressions with linking function – e.g. *said in other words* (*jinými slovy*). The problem here is that the boundaries between fixed and free combinations are not clear but they rather form a scale between these two. Therefore, it could be problematic to clearly state whether a certain expression is already fixed or not. Then it could happen that two similar expressions would be interpreted differently (one as a connective, another as an AltLex). From these reasons, we would prefer to classify these expressions as AltLex's (as they

are treated in PDTB) to avoid the decision on their fixedness. Another argument for this is that such expressions are annotated rather inconsistently (see the discussion on the expressions with *speaking* above). However, this issue needs a further discussion.

3.1.1 Further syntactic characterization

In the next step, we have examined the found AltLex's in terms of their integration in the clause structure, i.e. whether the experssion is an element modifying another element or a whole clause (as a clause modifier). Then we examined the syntactic structure of the Czech AltLex's in order to find out whether these expressions prefer some structures and follow certain patterns or not.

The analysis demonstrated that 78 types of AltLex's (83 %) are integrated in the clause structure and fulfill certain role of a clause element whereas 16 types (17 %) do not. These 16 types either comment the whole clause (as the so called disjuncts) or serves only as text-structuring expressions that do not contribute to the content of the clause – see several instances in Table 3:

	AltLex's		Total
	Integrated in the Clause Structure	**Non-Integrated in the Clause Structure**	
Examples	*Different (jiný)*	*To understand (rozumějme)*	
	Because of that (kvůli tomu)	*Translated (přeloženo)*	
	In the same breath (stejným dechem)	*As seen (jak je vidět)*	
	Similarly (podobně)	*The truth is (pravda je)*	
	Despite these facts (i přes tato fakta)	*Simply speaking (jednoduše řečeno)*	
	The consequence of this step (důsledkem tohoto kroku)		
	This is the reason why (to je důvod, proč)		
Number of AltLex types	78	16	94
%	83	17	100

Table 3: Syntactic characterization of Czech AltLex's: Integration in the clause structure

Another parameter examined the Czech AltLex's in terms of syntactic phrases. The analysis demonstrated that these expressions are realized either by noun phrases (NP), adjectival phrases (AdjP), numeral phrases (NumP), verbal phrases (VP), adverbial phrases (AdvP), prepositional phrases (PrepP), particle phrases (PartP) or by a whole clause.

3.1.1.1 Prepositional phrases

The largest group appeared to be prepositional phrases with the 33 types of AltLex's. Among them, it two subclasses have emerged. The first contains expressions where semantics and the property of being AltLex are

carried by the preposition. These AleLex's consist of a secondary preposition and an anaphoric expression that may vary.. The example is *in conflict with this* (*v rozporu s tím*). The fixed part carrying the meaning and signaling the type of a discourse relation is *in conflict with* (*v rozporu s*) that is classified as a secondary preposition in Czech. The second part is an anaphoric expression *this* (*tím*) that may vary (*in conflict with this/these facts/what was said* etc.).

The second subclass includes such expressions that are formed by a primary prepositions and a fixed noun signaling that it is an AltLex and indicating the type of the discourse relation – e.g. *from this reason* (*z tohoto důvodu*). In this example, it is the word *reason* indicating that there is a relation of reason – result.

3.1.1.2 AltLex's functioning as a whole clause

The second largest group contains the Czech AltLex's realized by a whole clause. Again they crystallized into two subclasses. The first (and larger) are clauses containing a semantically weak verb (e.g. *be*, *make*, *give*, *serve*) and the core meaning is carried by a noun, adjective or adverb – e.g. *the reason is* (*důvodem je*), *the difference is* (*rozdílem je*), *the exception is made* (*výjimku tvoří*), *it serves as an example* (*jako příklad slouží*), *he gives as the reason that* (*jako důvod uvádí, že*). The task, therefore, is whether it is better to treat them as whole clauses or to re-classify them under nominal, adjectival or adverbial phrases according to their semantically most relevant elements.

The second subclass of the whole clauses are those containing a non-finite verb (infinitive or participle) – e.g. *simply speaking* (*jednoduše řečeno*), *as seen* (*jak je vidět*), *it is necessary to add* (*dlužno dodat*). All of them function as disjuncts, i.e. clause modifiers. These expressions were treated as a whole clause, as they appear in a fixed clausal form and the simple verbs themselves do not function as AltLex's – e.g. the verbs *to speak* or *to see* do not signal any discourse relation on their own. It means they are not AltLex's inherently but only in the connection with another expressions with which they collocate. On the other hand, their head or core is formed by the verb, which may be an argument to treat them under the verbal phrases. In that case, it would be possible to cancel the group "whole clauses" completely, as all the present tokens may be placed under something else. However, each of the two present subclasses demonstrates some specific features and, therefore, we have left them together for the first phaze of research.

3.1.1.3 Verbal phrases

The third largest group contains verbal phrases. The heads of them are verbs that themselves signal a certain type of discourse relation and do not have to combine with other expressions to become an AltLex (as, e.g., the type *as seen* discussed above).They are lexically free, which means that they may occur in their whole paradigm and are not restricted to a limited set of forms. The examples are the verbs like *precede* (*předcházet*), *follow* (*následovat*), *give reasons* (*zdůvodnit*) – cf.:

(3)   *Gyla Horn agrees with the possible establishing of the property tax.*
He <u>gave the reason</u> that tightening of belts cannot be applied only to people living on wages.

Gyula Horn se vyslovil pro možné zavedení majetkové daně.
<u>*Zdůvodnil*</u> *to tím, že utahování opasků se nemůže vztahovat pouze na lidi žijící ze mzdy.*

The exact numbers of AltLex types for the individual syntactic phrases are given in Table 4[2]:

---

[2] The classification is done only for Czech, not for English; therefore, the English counterparts do not have to correspond to all of the given syntactic phrases.

		Examples	Number of AltLex types	%
Noun Phrases		*In the same breath (stejným dechem)*	2	2
		A while later (chvilku nato)		
Adjectival Phrases		*Other (další)*	2	2
		Different (jiný)		
Numeral Phrases		*The first – the second... (první – druhý...)*	1	1
Verbal Phrases		*Precede (předcházet)*	19	20.3
		Follow (následovat)		
		Give reasons (zdůvodnit)		
		Cause (způsobit)		
		Contrast (kontrastovat)		
Adverbial Phrases		*Later (později)*	4	4.3
		Precisely (přesněji)		
		Initially (původně)		
		Simultaneously (současně)		
Prepositional Phrases	**Subclass 1**	*In conflict with this (v rozporu s tím)*	18	19
		Because of that (kvůli tomu)		
		Not speaking of (nemluvě o)		
		Unlike that (na rozdíl od toho)		
	Subclass 2	*From this reason (z tohoto důvodu)*	15	16
		In the consequence (v jehož důsledku)		
		In this connection (v této souvislosti)		
		For this purpose (pro tento účel)		
Particle Phrases		*Truth (pravda)*	6	6.4
		What's more (tím spíš)		
		Just the same (právě tak)		
Whole Clause	**Subclass 1**	*The reason is (důvodem je)*	19	20
		The exception is (výjimkou je)		
		The result is (výsledkem je)		
		He gives an example (jako příklad uvedl)		
	Subclass 2	*To understand (rozumějme)*	8	9
		Translated (přeloženo)		
		As seen (jak je vidět)		
		Simply speaking (jednoduše řečeno)		
Total			94	100

Table 4: Syntactic characterization of Czech AltLex's: Syntactic Phrases

3.1.2 Further lexical characterization

If we look at the Czech AltLex's from the lexical aspect, we find out that they form a scale with two polar ends. The first represents expressions containing a word that is AltLex inherently, i.e. it signals a certain discourse relation on its own and forms several open collocations (or free combinations), i.e. an open sequence with no mutual expectancy that is grammatically and lexically unrestricted. An illustration are, for example, verbal AltLex's that may use their whole paradigm, i.e. they may occur in all tenses, both in active and passive form, with modal expression etc. – cf. the found instances of one AltLex type, the verb *to add* (in the sense of *saying as a further remark*) : *it is necessary to add* (*k tomu je třeba dodat*), *he added* (*dodal*), *a member of the organization adds* (*dodává člen organizace*), *we should add* (*dodejme*).

The second pole includes multiword expressions whose items become an AltLex only in a particular combination and are both lexically and grammatically restricted. They allow only a slight modification (i.e. they occur in a limited set of variants like *simply/shortly/generally speaking – jednoduše/krátce/obecně řečeno*) or they are fully frozen (restricted to a single combination – e.g. *what's more – o to více*). Usually, such expressions are not complete grammatical structures. These AltLex's are text oriented lexical bundles characterized as "the most frequently recurring lexical sequences" (Biber and Conrad, 1999: 183) that participate in organization and structuring of the text.

However, not all of the examined expressions allowed for an exact categorization (i.e. it was impossible to state whether they are lexically free or fully frozen). E.g. the AltLex *serve as an example* (*sloužit jako příklad*) is not frozen (it is not an incomplete grammatical structure and the verb may be conjugated) but, at the same time, it exhibits a certain degree of expectancy and predictability, which is typical for fixed expressions. Therefore, we avoid the strict categorization of either/or and we understand the AltLex's as a scale or continuum from fully free to fully frozen combinations. At the same time, it is necessary to point out that the frozen expressions are in minority and that the majority of Czech AltLex's occur toward the free combination pole.

3.2   Semantic characterization

Within other cohesive devices, discourse markers have a special two-part position. They signal a discourse relation and contain an anaphoric expression that refers to the first argument (Forbes-Riley et al., 2006). At the same time, the anaforic reference may be either explicit or implicit (Prasad et al. gives an example of two expressions *as a result of that* and *as a result* that have the same meaning). Prasad et al. argue that English AltLex's have the same although more complex two-part semantic contribution.

The situation in Czech seems to be very similar. Czech AltLex's also include an anaphoric reference that may be explicit or implicit. For some of them, to express or not to express an anaphoric reference is even obligatory – see Table 5. The category "obligatory implicit" means that the expressions have no possibility to express the anaphoric reference on the surface layer – cf. it is impossible to say *this simply speaking, ...* (*toto stručně řečeno, ...*) but only *simply speaking, ...* (*stručně řečeno, ...*). The class "obligatory explicit" contains expressions that are ungrammatical without the anaphoric reference – e.g. we cannot say *because of* (*kvůli*) but only *because of that* (*kvůli tomu*). The category of "optional" anaphoric reference means that the expressions have two possibilities – either to express the reference on the surface (i.e. explicitly) or not (i.e., they express it only implicitly)

		Obligatory	Optional
Implicit	*Simply speaking (jednoruše řečeno)*	*He added (dodal)*
	Translated (přeloženo)	*The first – the second... (první – druhý...)*
	As seen (jak je vidět)	*The reason is (důvodem je)*
	In the same breath (stejným dechem)	*An example is (příkladem je)*
Explicit	*From this reason (z tohoto důvodu)*	*The reason of this is (důvodem toho je)*
	Because of that (kvůli tomu)	*The consequence of this step is (důsledkem tohoto kroku je)*
	Despite these facts (i přes tato fakta)	*(It) is connected with (souvisí to s)*
	This is in contrast (s tím kontrastuje)	*(It) is not valid in the case (neplatí to v případě)*

Table 5: Implicit and explicit anaphoric reference – instances

The analysis demonstrated that AltLex's with obligatory implicit reference are lexically frozen expressions that do not allow free combinations with other words, i.e. with an anaphoric reference as well (*translated – přeloženo*). Another group of AltLex's includes expressions that express the anaphoric reference obligatory. These are verbs that do so because of their valency – e.g. the verb *to contrast* (*kontrastovat*) requires a complementation of patient that is, in the case of AltLex, anaphoric; therefore, it is impossible to say *another fact contrasts* (*jiná skutečnost kontrastuje*), but *another fact contrast with this* (*s tím kontrastuje jiná skutečnost*). Anaphoric reference is expressed obligatory also by the AltLex's whose head is a preposition requiring the complementation of an

anaphoric reference like *because of that* (*kvůli tomu*), *despite these facts* (*i přes tato fakta*) etc.

The AltLex's with the optional anaphoric reference are partly the same as in English. These are the expressions of the type *the result (of this) is* (*výsledkem /toho/ je*) and with the ellipsis of a noun in expressions like *the second (step) is* (*druhým /krokem/ je*). In addition to them, there is another class of AltLex's determined by language resulting from the fact that Czech allows a surface omission of subject. Therefore, if the potential anaphoric reference occurs in the position of subject, it may be omitted – cf. the instances like *(it) is connected with* (*souvisí to s*), *(it) is not valid in the case* (*neplatí to v případě*) etc.

Out of the total number (94) of AltLex types, 41 % express the anaphoric reference optionally, 31 % obligatory and 28 % cannot express it on the surface layer at all. This demonstrates that the three possibilities are rather balanced – see Table 6 (Optional types are not divided into implicit and explicit because they have a possibility of both. The expression of an anaphoric reference in the surface depends on the individual tokens, not on the AltLex type.):

Types of AltLex's (out of 94)		Number		%	
Optional types (exist in boht implicit and explicit variants)		39			41
Obligatory types	Implicit	26	55	28	59
	Explicit	29		31	
Total		94			100

Table 6: Implicit and explicit anaphoric reference – types of AltLex's

In addition to AltLex types, we have also counted their actual tokens in PDT in order to find out whether the AltLex's with the optional anaphoric reference prefer to express it or not. It means, for example, whether there is a tendency to say *an example of this is* (*příkladem toho je*) or *an example is* (*příkladem je*). See Table 7:

The actual tokens (out of 261)	Obligatory	Optional	Total
Implicit	35	98	133
Explicit	60	68	128
Total	95	166	261

Table 7: Implicit and explicit anaphoric reference – the actual tokens of AltLex's

The analysis demonstrated that PDT contains 164 AltLex's expressing the anaphoric reference optionally. Out of this number, 98 instances (59 %) appeared with the expressed reference and 68 (41 %) without it. Therefore, it seems that if the AltLex has a possibility of choice, there is a slight tendency not to express the anaphoric reference. However, the present number of AltLex's in PDT is not final. Therefore, we should treat this observation as a hypothesis that is necessary to verify on a larger amount of data.

4.   Conclusion

The present annotation of the upcoming PDT version demonstrates that many discourse relations in Czech are not realized with the "classic" connectives but by other means – by the so called alternative lexicalizations of discourse markers. Without taking these expressions into account, the annotation of discourse would be much poorer and incomplete. In terms of lexico-syntactic and semantic characterization, Czech AltLex's seem to exhibit similar features as their English counterparts with some exceptions that are determined by the different language type. The analysis above demonstrated that the current stage of AltLex annotation in PDT is in the beginning and needs further re-annotation intended for the next PDT version.

5.   Acknowledgements

6.   References

Biber, D. and Conrad, S. (1999). Lexical bundles in conversation and academic prose. In *Out of Corpora: Studies in Honour of Stig Johansson*, ed. by Hasselgard, H. and Oksefjell, S. Amsterdam: Rodopi.

Forbes-Riley, K., Webber, B. and Joshi, A. (2006). Computing discourse semantics: The predicate-argument semantics of discourse connectives in D-LTAG. *Journal of Semantics*, 23:55–106.

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

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

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

Mladová, L., Zikánová, Š., Bedřichová, Z., Mírovský, J., Jínová, P., Zdeňková, J., Rysová, M., Hajičová, E.

(2011). *Manual for Annotation of Textual Relations (of Discourse) in the Prague Dependency Treebank.* Praha: UFAL MFF. To appear.

Prasad, R. et al. (2010). *Realization of Discourse Relations by Other Means: Alternative Lexicalizations.* In COLING (Posters).

6.1   Data sources

Hajič, J. et al. (2006). *Prague Dependency Treebank 2.0.* Philadelphia: Linguistic Data Consortium.