On Discourse Annotation in PDT

Magdaléna Rysová
magdalena.rysova@ufal.mff.cuni.cz
Charles University in Prague,
Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics

I) Discourse connectives vs. their alternative lexicalizations (= AltLex’s)

II) Their current annotation in the Prague Dependency Treebank (PDT)

III) Possible future annotation of Czech AltLex’s in PDT

IV) Lexico-syntactic and semantic characterization of Czech AltLex’s

V) Patterns of Czech AltLex’s

VI) Lemmatization of Czech AltLex’s

VII) Possible future issues (what we need to solve)

I) Discourse connectives vs. their alternative lexicalizations (= AltLex's)

I) Discourse connectives vs. their alternative lexicalizations (= AltLex's)

Connectives = expressions with connecting function at the level of discourse description

a) Coordinating conjunctions: *and* (*a*), *but* (*ale*), *therefore* (*proto*);

b) subordinating conjunctions: *although* (*ačkoliv*);

c) particle expressions (including rhematizers): *even (dokonce), too (také);*

d) adverbs: *then* (*potom*);

e) certain uses of pronouns: *except for this* (*kromě toho*);

f) idiomatic multiple-word connective means formed by linking of different expressions: *on the one hand* (*na jedné straně*);

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

h) two punctuation marks: colon (:) and dash (–).

Expressions with the same function but from other classes = **alternative lexicalizations of discourse connectives (= AltLex's)**

I) Discourse connectives vs. their alternative lexicalizations (= AltLex's)

connective vs. AltLex

=

***therefore* (*proto*) vs. *the reason is* (*důvodem je*)**

Other examples of AtLex's in PDT:

upřesnit (specify)
být výsledkem (be the result)
s odůvodněním (with justification)
souviset (be related)
vyplývat (entail)
za tím účelem (for that purpose)
důsledkem tohoto kroku (the result of that step)
způsobit (cause)

II) Their current annotation in the Prague Dependency Treebank (PDT)

II) Their current annotation in the Prague Dependency Treebank (PDT)

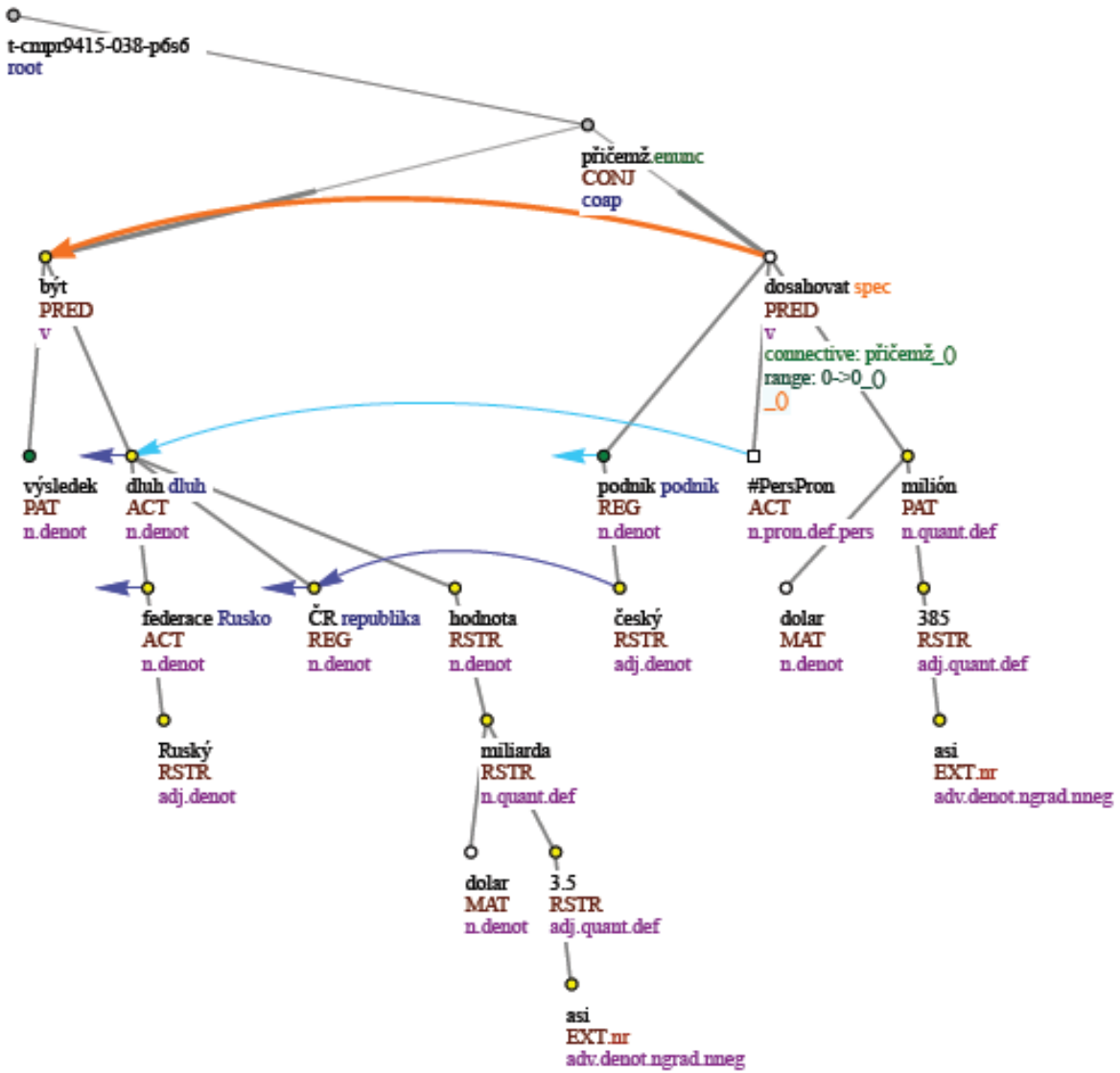
In the current stage of annotations, discourse relations are captured only if signaled by explicit discourse connectives (not AltLex’s).

Example of a connective from PDT:

*The result is a debt about $ 3.5 billion of the Russian Federation to the Czech Republic, **which** is about $ 385 million for Czech enterprises.*

*Výsledkem je dluh Ruské federace vůči ČR v hodnotě asi 3,5 miliardy dolarů, **přičemž** vůči českým podnikům dosahuje asi 385 milionů dolarů*.

II) Their current annotation in the Prague Dependency Treebank (PDT)



II) Their current annotation in the Prague Dependency Treebank (PDT)

- In the current stage of annotations, discourse relations signaled by AltLex's are not captured.
- Such expressions are only marked with the annotator's comment "altlex".

II) Their current annotation in the Prague Dependency Treebank (PDT)

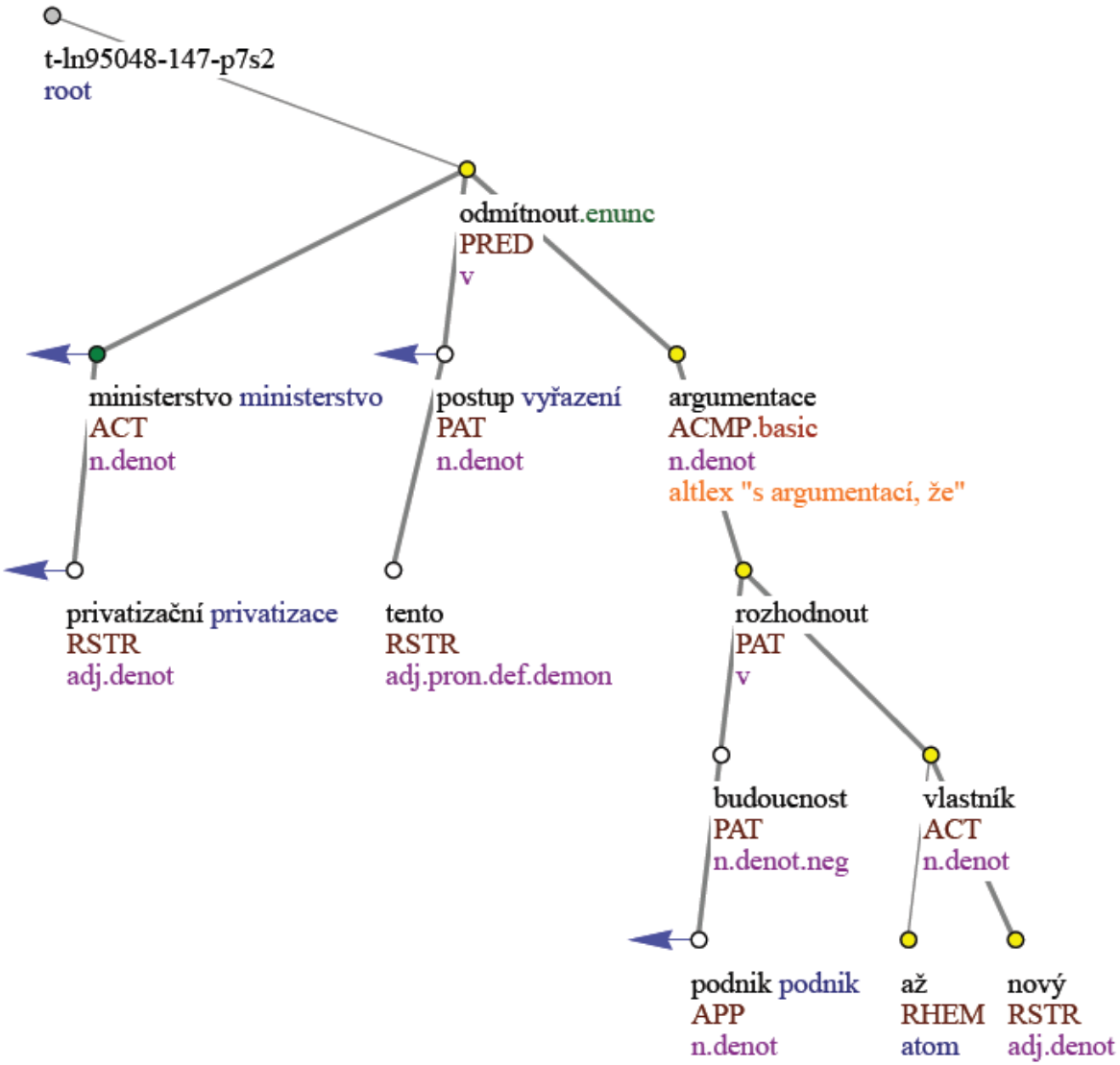
Example of an AltLex from PDT:

Privatization Ministry rejected this approach ***with the argument that*** *the future of business should be decided by new owners.*

(Privatizační ministerstvo tento postup odmítlo ***s argumentací, že*** *o budoucnosti podniků by měli rozhodnout až noví vlastníci.)*

AltLex = ***with the argument that***

II) Their current annotation in the Prague Dependency Treebank (PDT)



III) Possible future annotation of Czech AltLex's in PDT

III) Possible future annotation of Czech AltLex's in PDT

- Both connectives and AltLex's have similar function – to connect two discourse arguments.
- Therefore, there is no need to capture only the relations signaled by one of them.
- The task for the next stage – to annotate discourse relations expressed by AltLex's.

III) Possible future annotation of Czech AltLex's in PDT

Example from PDT – interchangeability of connectives and AltLex's:

*The Brazilian football player attacked his opponent in today's match. **This is the reason why** he will not play in the next three matches.*

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

*(Hráč brazilského týmu napadl v dnešním utkání svého protihráče. **To je důvod, proč / Proto** nebude hrát příští tři zápasy.)*

III) Possible future annotation of Czech AltLex's in PDT

Another reason why to annotate AltLex's:

- Some cases are captured on the tectogrammatical layer rather according to their structure than to their meaning.
- Preposition *with* + noun – *with argumentation* (*s argumentací*), *with condition* (*s podmínkou*)
- Tectogrammatical layer: functor ACMP

III) Possible future annotation of Czech AltLex's in PDT

Functor ACMP – basic forms (cf. *Annotation on the tectogrammatical level in the Prague Dependency Treebank*, 2006):

- **prepositional phrases –** *He works without glasses (Pracuje bez brýlí.); He walks with a stick (Chodí s holí.)*

- **dependent clauses** – e.g. *with the fact that* (*s tím, že)* – *They bought two sets of lego, planning to give one to each of their sons; lit. with the fact that they give... (Koupili dvě sady lega s tím, že dají každému synovi jednu.)*

III) Possible future annotation of Czech AltLex's in PDT

- Functor ACMP – any accompanying circumstance; the semantics is not further specified

→ no distinction between:

He came with a stick. (Přišel s holí.) and *He came with the justification that…* (*Přišel s odůvodněním.*)

III) Possible future annotation of Czech AltLex's in PDT

- There is a large group of AltLex's with the functor ACMP on the tectogrammatical layer

- Structure: *with* + noun (*argument, justification, condition*…) + *that*

- Future annotation of AltLex's also based on their structure.

III) Possible future annotation of Czech AltLex's in PDT

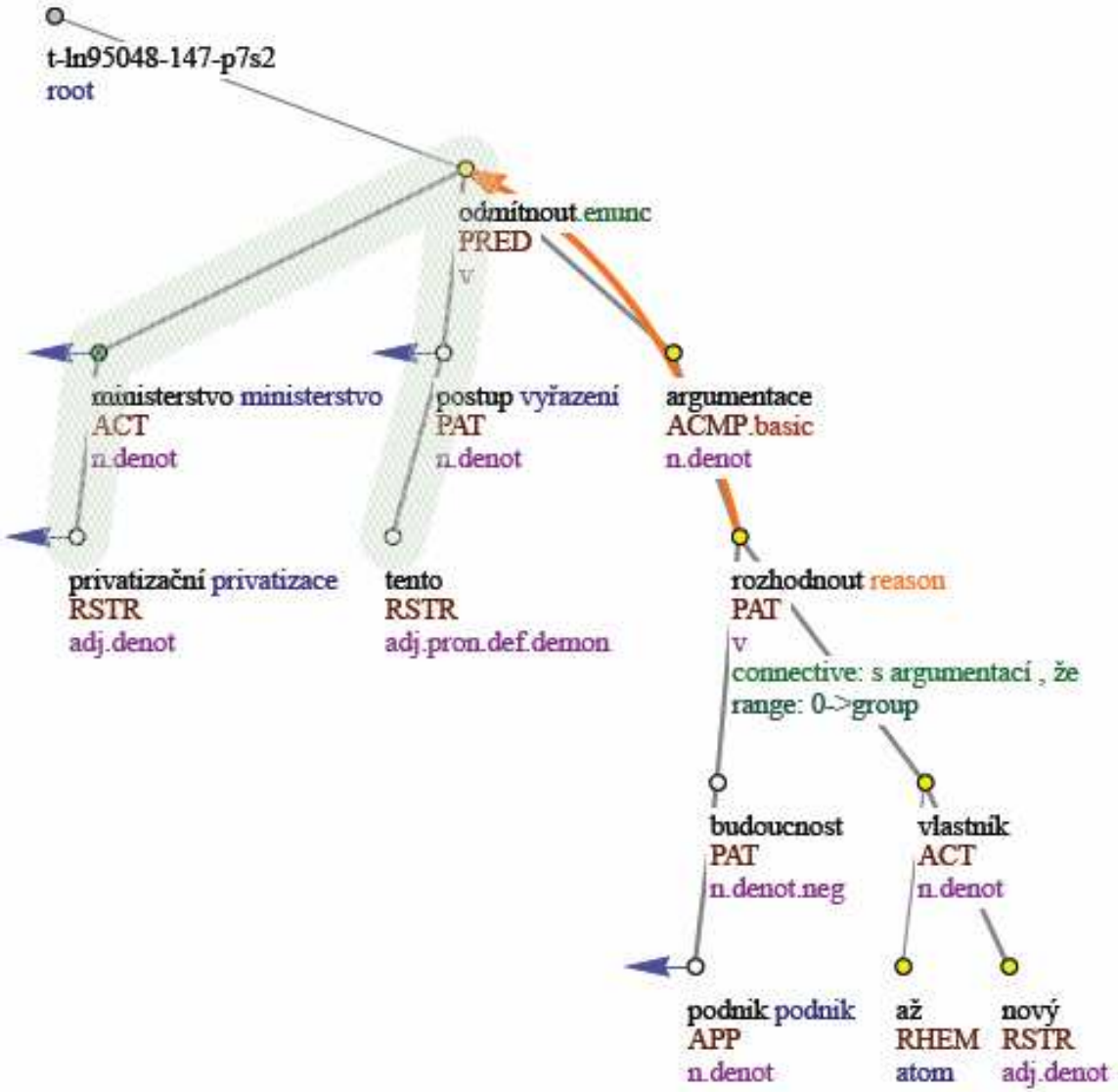
Example from PDT:

Privatization Ministry rejected this approach ***with the argument that*** *the future of business should be decided by new owners.*

(*Privatizační ministerstvo tento postup odmítlo* ***s argumentací, že*** *o budoucnosti podniků by měli rozhodnout až noví vlastníci.*)

AltLex = ***with the argument that***; discourse relation = reason – result

III) Possible future annotation of Czech AltLex's in PDT



III) Possible future annotation of Czech AltLex's in PDT

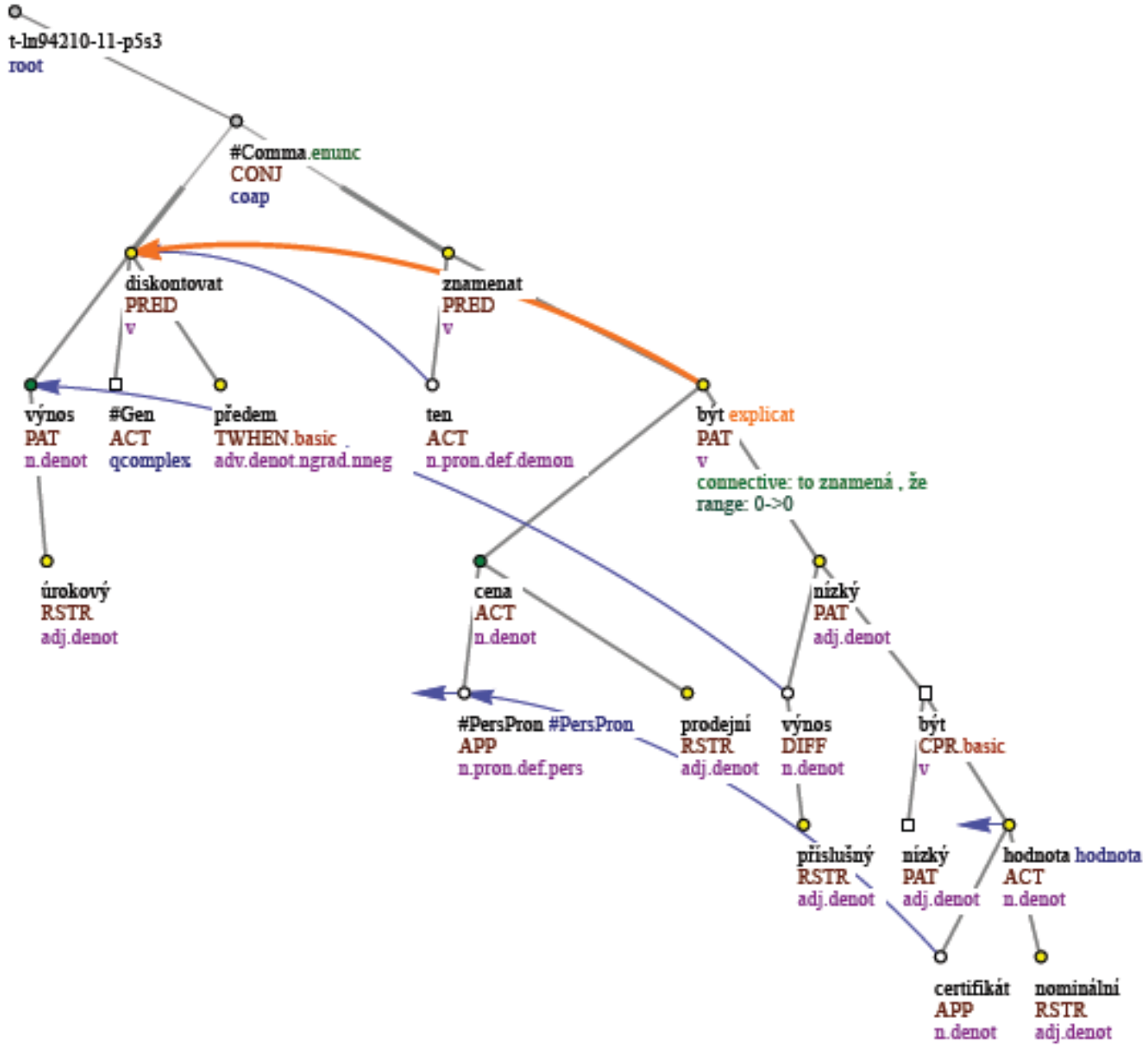
Example from PDT:

*Interest income is discounted in advance, **which means that** its selling price is lower by the revenue than the nominal value of the certificate.*

*(Úrokový výnos je diskontován předem, **to znamená, že** jeho prodejní cena je o příslušný výnos nižší než nominální hodnota certifikátu.)*

AltLex = *which means that*; discourse relation = explication

III) Possible future annotation of Czech AltLex's in PDT



IV) Lexico-syntactic and semantic characterization of Czech AltLex's

IV) Lexico-syntactic and semantic characterization of Czech AltLex's

- All expressions within the total 43,955 of sentences in PDT that were annotated as AltLex's.
- **261 tokens** of Czech AltLex's and **94 AltLex types** (e.g. the type *the reason is* appeared in 17 tokens)

IV) Lexico-syntactic and semantic characterization of Czech AltLex's

Inconsistencies in annotation of AltLex's in PDT:

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

IV) Lexico-syntactic and semantic characterization of Czech AltLex's

1) Syntactic characterization

IV) Lexico-syntactic and semantic characterization of Czech AltLex's

1.a Integration in the clause structure:

AltLex's	Types	%
Integrated in the Clause Structure	78	83
Non-integrated in the Clause Structure	16	17
TOTAL types	94	100
TOTAL occurrences	261	

Examples:

- integrated: *different – jiný*, *similarly – podobně*
- non-integrated (disjuncts): *as seen – jak je vidět*

IV) Lexico-syntactic and semantic characterization of Czech AltLex's

1.b Syntactic structure

- noun phrases, adjectival phrases, numeral phrases, verbal phrases, adverbial phrases, prepositional phrases, particle phrases or whole clauses

Most frequent:

- prepositional phrases (e.g. *in conflict with this – v rozporu s tím*)
- whole clauses (e.g. *the difference is – rozdílem je*)
- verbal phrases (e.g. *precede – předcházet*)

IV) Lexico-syntactic and semantic characterization of Czech AltLex's

1) Prepositional phrases (33 types of AltLex's)

a) secondary preposition + an anaphoric expression (*in conflict with this/these facts/what was said – v rozporu s tím/těmito fakty/s tím, co bylo řečeno)*

b) primary preposition + a fixed noun signaling that it is an AltLex (*from this reason – z tohoto důvodu*)

IV) Lexico-syntactic and semantic characterization of Czech AltLex's

2) Whole clauses (27 types of AltLex's)

a) semantically weak verb (e.g. *be*, *make*, *give*, *serve*) + a noun, adjective or adverb carrying the core meaning – e.g. *the reason is* (*důvodem je*), *the difference is* (*rozdílem je*)

b) non-finite verb (infinitive or participle) – e.g. *it is necessary to add* (*dlužno dodat*), *as seen* (*jak je vidět*)

IV) Lexico-syntactic and semantic characterization of Czech AltLex's

3) Verbal phrases (19 types of AltLex's)

- The heads are verbs that themselves signal a certain type of discourse relation and do not have to combine with other expressions to become an AltLex

- Lexically free – they may occur in their whole paradigm and are not restricted to a limited set of forms – *precede* (*předcházet*), *follow* (*následovat*), *give reasons* (*zdůvodnit*)

IV) Lexico-syntactic and semantic characterization of Czech AltLex's

2) Lexical characterization

IV) Lexico-syntactic and semantic characterization of Czech AltLex's

a) Expressions containing **a word that is AltLex by itself** (forming several open collocations with no mutual expectancy that is grammatically and lexically unrestricted):

- *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*

IV) Lexico-syntactic and semantic characterization of Czech AltLex's

b) Multiword expressions whose items become an AltLex only in a particular combination and are both lexically and grammatically restricted:

- *simply/shortly/generally speaking – jednoduše/krátce/obecně řečeno*

IV) Lexico-syntactic and semantic characterization of Czech AltLex's

3) Semantic characterization

IV) Lexico-syntactic and semantic characterization of Czech AltLex's

Czech AltLex's

a) signal certain discourse relation;

b) contain an anaphoric expression that refers to the first argument:

IV) Lexico-syntactic and semantic characterization of Czech AltLex's

i) anaphoric reference **may be** expressed on the surface layer:

- *an example of this is* (*příkladem toho je*) vs. *an example is* (*příkladem je*)

ii) anaphoric reference **must be** expressed on the surface layer:

- *another fact contrast with this* (*s tím kontrastuje jiná skutečnost*); not **another fact contrasts* (**jiná skutečnost kontrastuje*)

iii) anaphoric reference **cannot be** expressed on the surface layer:

- *simply speaking, …* (*stručně řečeno, …*); not **this simply speaking, …* (**toto stručně řečeno, …*)

V) Patterns of Czech AltLex's

(based on the patterns of English AltLex's – Prasad et al. 2010)

V) Patterns of Czech AltLex’s

A group of Czech AltLex’s (forming several open collocations with no mutual expectancy that is grammatically and lexically unrestricted) may be analyzed in terms of their **general patterns** according to which it is possible to generate their **concrete realizations** in texts.

V) Patterns of Czech AltLex's

These AltLex's have their **core** – a word signaling a certain discourse relation that appears in several open colocations:

- *reason* (*důvod*): *the reason is* (*důvodem je*), *that is the reason why* (*to je důvod proč*)…

- *consequence (důsledek): a consequence was* (*důsledkem bylo*), *a main consequence of this is* (*hlavním důsledkem toho je*)…

Reason/justification (důvod)

- 6 types of Czech AltLex’s (within 94)
- A general pattern for each of them
- Prep – preposition, Adj – adjective, SubC – subordinate clause, MainC – main clause, VP – verbal phrase, NP – nominal phrase, anaph. – anaphoric reference; facultative modifications marked with round brackets

V) Patterns of Czech AltLex’s

1) <Prep> (<Adj>) *odůvodnění* <SubC>

- i.e. *with (later) justification that –*

s (pozdějším) odůvodněním, že

2) <Prep> <anaph. Pron>/<anaph. Adj>/<anaph. NP> *důvod* <MainC>

- i.e. *from this/the introduced reason / from the reason of this fact – z tohoto/uvedeného důvodu / z důvodu této skutečnosti*

V) Patterns of Czech AltLex’s

3) <anaph. Pron>/<anaph. NP> <VP> (<Adj>) *důvod* <NP>/<SubC>

- i.e. *this/this fact is (the main) reason why/of the success – to / tato skutečnost je (hlavní) důvod, proč / úspěchu*

4) (<Adj>) *důvod* (<anaph. Pron>/<anaph. NP>) <VP> <NP>/<SubC>

- i.e. *(the main) reason (of this/the division) is that / the fact – (hlavním) důvodem (toho/rozdělení) je, že / skutečnost*

V) Patterns of Czech AltLex’s

5) <VP> (<Prep>) (<Adj>) *důvod* (<anaph. Pron>/<anaph. NP>) <NP>/<SubC>

- i.e. *to give (as) (the main) reason (of this/the departure) that / the fact – uvést (jako) (hlavní) důvod (toho/odchodu), že / skutečnost*

6) *z/odůvodnit* <anaph. Pron>/<anaph. NP> <NP>/<SubC>

- i.e. *to justify it / this situation by the fact that / by the complexity – z/odůvodnit to / tuto skutečnost tím, že / složitostí*

V) Patterns of Czech AltLex's

- It is possible to form similar patters for other types of AltLex's except for lexically frozen or restricted expressions – e.g. *as seen* (*jak je vidět*).
- According to the core words, it will be possible to find other concrete realizations of the individual AltLex's in PDT.

VI) Lemmatization of Czech AltLex’s

Lemmatization of Czech AltLex's

1) AltLex's with a **key word** – *reason* (*důvod*), *consequence* (*důsledek*), *condition* (*podmínka*), *example* (*příklad*), *case* (*případ*)

- forming various open colocations

2) Expressions functioning as AltLex's only in **combination with an anaphoric expression** that refers to the previous argument – prepositions like *because of* (*kvůli*), *unlike* (*na rozdíl od*), *due to* (*vinou*), *despite* (*navzdory*)

Lemmatization of Czech AltLex's

Italy saves. Because of this, some diaries will no longer come out.

(Itálie šetří. Kvůli tomu tam přestanou vycházet některé deníky.) – anaphoric, is AltLex

vs.

I was ill a whole month. I could not sleep due to cough at night.

(Marodila jsem celý měsíc. V noci jsem nemohla spát kvůli kašli.) – not anaphoric, is not AltLex

3) Lexically frozen or restricted expressions
– *translated* (*přeloženo*), *to understand* (*rozumějme*), *as seen* (*jak je vidět*)

Lemmatization of Czech AltLex’s

A list of the found AltLex’s in PDT	
argumentovat	přeloženo
další	přes
díky	přesněji
dlužno dodat	přičíst
do třetice	příčina
dodat/dodávat	přidávat
doplňovat	příklad: být příkladem; příklad pochází z; jako příklad slouží; uvést jako příklad
důsledek: důsledkem tohoto kroku je; v jehož důsledku; to je důsledek	případ: být opačným případem; neplatí to v případě, v (tomto) případě; to je případ; v horším případě
důvod: s odůvodněním; z tohoto důvodu; z těchto důvodů; z uvedených důvodů; to je důvod, proč; důvodem je; důvodů je několik; uvést jako důvod	přispívat
jak je vidět	přistupovat

Lemmatization of Czech AltLex’s

jako poslední	rozdíl
jiný	rozpor
konečný	rozumějme
kontrastovat	řečeno: stručně řečeno; jinak řečeno; zjednodušeně řečeno; jednoduše řečeno
kvůli	se slovy, že
mezi	souviset
na prvním místě	stejným dechem
na rozdíl od	tím spíš
na základě	týkat se
na závěr	účel: za tím/tímto účelem; pro tento účel
následovat	upřesnit

Lemmatization of Czech AltLex’s

navzdory	v (této) souvislosti
nehledě na	ve skutečnosti
nemluvě o	vést
o to více	vinou
podmínka: být podmínkou; s podmínkou, že	výjimka: být výjimkou; tvořit výjimku
podobně	vyplývat
pokračovat	výsledek
pravda: pravda; pravda je	vzhledem k
pravděpodobnější	zdůvodnit/zdůvodňovat/odůvodnit/odůvodňovat
právě tak	znamenat
předcházet	způsobit

VII) Possible future issues (what we need to solve)

VII) Possible future issues (what we need to solve)

1. Terminology:

- Is “alternative lexicalization of discourse connectives” a suitable term?
- There are such discourse relations that lack a connective and are expressed only by an AltLex (English: no adverbial connective for “Cause:Reason” – only AltLex’s like “a major reason is” – cf. Prasad et al. 2010)
- Therefore, is the word “alternative” suitable?

VII) Possible future issues (what we need to solve)

2. Do we need to have two terms for expressions indicating discourse relations?

Why to have two terms (connectives and AltLex's) if they have similar functions?

Would it not be more useful to have only one umbrella term "discourse connectives" for both "classic" connectives and AltLex's? (cf. Hoffmannová 1993)

References

- Hoffmannová, J. *Sémantické a pragmatické aspekty koherence textu*. Praha: Ústav pro jazyk český, 1983.

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

Data sources

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

Thank you for your attention.

Acknowledgement

This talk was supported by the following grants and projects:

P406/10/0875 Computational Linguistics: Explicit description of language and annotated data focused on Czech (Czech Science Foundation)

P406/12/0658 Coreference, discourse relations and information structure in a contrastive perspective (Czech Science Foundation)

ME10018 Towards a computational analysis of text structure (Ministry of Education, Youth and Sports of the CR /MSMT/)