**Czech UMR data, version 1.0 – notes**

**I. Source Data**:

* PDT-C .. "almost 2.0" version
  + how big data – total numbers : 175 429 sentences
  + problem – data not publically released (where to find them?)
* format conversion (PDT -> Treex). In Treex, there are 3 stages:

1. Building the U-tree from the tectogrammatical tree
2. Process the coreference
3. Adjust special structures (coordination, CONTRD)

**II. Sentence Level Representation**

**II.1 Structural changes**

* **coordination** … change of the structure to be in compliance with the UMR specification (discourse relations; reification in 2 subtypes)
  + coordination … POZOR, někde víc ARG než 2 (cca 30) případů
    - "but-91", ??? kde ještě ???
* **apposition** … as identity-91

většinou dvoučlenné, ale v PDT-C 1.0 je:

- několik málo desítek apozic se 3 členy

- pár apozic se 4 členy

- 4x apozice s 5 členy ... 1x nějaký nesrozumitelný výčet, 3x chyba (viz níž)

**Vycházím z manuálového příkladu:** *Přijeli do měst, jako.APPS (Praha, Brno a Ostrava).CONJ*

* + - "... a pak se to musí ochutit :.CONJ sůl , pepř , nové koření , majoránka ,.APPS česnek ."

Nemělo by to být opačně, apozice "ochutit" a to koření (které je koordinované), tedy "... a pak se to musí ochutit :.APPS sůl , pepř , nové koření , majoránka ,.CONJ česnek??

* + - "... si tam člověk udělá pohodlí ,.APPS může si v klidu číst , poslouchat muziku , najíst se , napít se "

OK apozice mezi "udělat si pohodlí" a 4 činnostmi "číst", "poslouchat", "najíst se", "napít se" ... ale nejsou ty 4 činnosti spíš koordinované (DISJ)?

* + - "... objevují či znovuobjevují texty emigrantů jako.APPS Ivana Bunina, Vladimira Nabokova, Fridricha Gorenštejna, Josifa Brodského, Sergeje Dovlatova ..."

OK apozice uvozená "jako", ... ale nejsou ty 4 činnosti spíš koordinované (asi CONJ)?

* **coreference-related structures** changed to be in compliance with the UMR specification:
  + re-entrances within a sentence, esp.:
    - anaphor is a personal or possessive pronoun (incl. reflexives): type *Maria – she – the girl*

but also with a nominal anaphor: type *Maria – the girl*

* + - * ??? hodnota gramatémů "Inher/inher " pro reflexiva (*se/si, svůj*) … zůstávají v datech (předpokládalo se, že zmizí??)
      * pro tuto chvíli DZ navrhuje smazat refer-number, refer-person
    - raising and control verbs

(UD: verbs with open clausal complement (xcomp) = verbs with predicative complement)

(incl. cases without overtly expressed anaphor, type *Martin viděl Petra přicházet*)

* + - keep separate nodes if they are further modified – and identify the coreference at the document level !!
    - #Rcp … coreference identified at the document level
  + inverse roles
    - relative clauses
      * ??? hodnota gramatémů "Inher/inher" pro relativa (*ti,* ***kdo/kteří*** *...*) … zůstávají v datech (předpokládalo se, že zmizí??)
      * pro tuto chvíli DZ navrhuje smazat refer-number, refer-person
  + ~~cílem koreferenčních šipek by měly být všechny koordinované členy (u APPS je cílem první z aponivaných členů) … viz níž~~
* **??? anything else ???**
  + **Problém: katafora …** UMR nerozlišuje kataforu a anaforu, šipka vede vždy do nějaké dřívější (nebo nanejvýš do aktuální) věty!!!

🡪 katafora způsobuje velké problémy (upozorňuje na ně validátor)

**Koreference kombinovaná s apozice a/nebo koordinací … zatím chybí!!!**

**Apozice a koreference:**

1. **šipka vede z uzlu APPS** (coref\_text/target\_node.rf ... 1551)

... přesunout ji tak, aby vedla z prvního z aponovaných členů (tj. o patro níž)??

V UMR koreference "same-entity" by měla být OK.

2. **šipka vede do uzlu APPS** (coref\_text/target\_node.rf i coref\_gram)

... přesunout ji tak, aby vedla do prvního z aponovaných členů (tj. o patro níž)??

V UMR koreference "same-entity" by měla být OK.

3. Platí to same, pokud dojde ke **kombinaci 1. a 2**.? (šipka vede mezi APPS uzly, 66 případů), tj. měly by se prolinkovat první členy apozice, relace "same-entity".

**Koordinace a koreference:**

1. **šipka vede z uzlu CONJ/DISJ/...**

... zmnožit šipku tak, aby šipky vedly ze všech členů koordinace (začátek-šipky) do původního cíle (konec-šipky)

V UMR koreference by neměla být "same-entity", ale "subset-of, a to takto:

**konec-šipky :subset-of začátek-šipky**

(tj. ty koordinované členy budou děti v relaci :subset-of )

2**. šipka vede do uzlu CONJ/DISJ/...**

... zmnožit šipku tak, aby šipky vedly z původního uzlu (začátek-šipky) do všech koordinovaných členů (konec-šipky)

V UMR koreference by neměla být "same-entity", ale "subset-o, a to takto:

**začátek-šipky :subset-of konec-šipky**

(tj. ty koordinované členy budou děti v relaci :subset-of )

3. Pokud dojde ke **kombinaci 1. a 2.** (šipka vede mezi spojovacími uzly, tj. uzly CONJ,/DISJ/…, 190 případů) … nejasné!!!

Možné řešení: Zatím zůstanou koreferenční šipky mezi celými koordinacemi (tedy mezi spojovacími uzly) tak, jak je to v PDT-C, se :same-entity relací.

(Tedy jinak než u jednoduchých koordinací, kde vede buď jen šipka do spojovacího uzlu, nebo jen šipka ze spojovacího uzlu.)

* + **Problém**: jaká relace … prozatím nechat **:same-entity**

**II.2 Nodes labeling … ??? hotovo**

**artificial t\_lemmas** (t\_lemma substitutes as, e.g., #Person) … "translated" to abstract concepts where appropriate/possible (or their supertypes whenever automatic disambiguation is not possible; e.g., "entity" subsumes both "person" and "thing")

* + !! supertypes:
    - entity (subsumes both "person" and "thing")
    - concept (subsumes "entity", "state", "event") in specific constructions, esp. constructions with the meaning of comparison (i.e., two or more events, states or entities are compared)
  + event rolesets: see II.3
* **POZOR: nejasné, co s qcomplex #Ast, #Period3, #Comma, #Colon, #Dash**

**II.3 Relations labeling**

**Verb-specific arguments labeling:**

* verb specific conversion for **43%** of verb predicates (= frames ≈ rolesets with arguments) (Hajič et al, 2024)
  + disambiguation needed(ca 25 frames with two possible mappings, mail JŠ, 9 May, 2024)
* the rest based on the PDT-Vallex lexicon
  + verb-specific conversion must be extended

**Default conversion table:**

* **SOME functors still missing … ??? hotovo**
* translation of some roles still too coarse (refinement needed)
* several **new labels** to cover PDT-specific annotation:
  + !! new **roles**
    - **effect** (EFF)
    - **comparison** (based on CPR, should be further inspected and refined)
    - **regard** (CRIT, REG)
  + !! new "**clausal-marker**" role
    - for rhematizers (RHEM),
    - sentence/ linking / modal adverbial expressions
      * attitude marker (ATT),
      * modal marker (MOD),
      * discourse marker (PREC),
      * conjunction modifier (CM)
  + !! new **discourse roles** 
    - **gradation** (GRAD)
    - **independent-clause**
      * for parentheses (PAR),
      * interjection (PARTL),
      * vocative clause (VOCAT)
  + !! new roles for **MWE** 
    - **predicative-noun** (CPHR) … tentative role used in light verb constructions, will be removed when LVCs are processed
    - **part-of-phraseme** (DPHR) … tentative role used for identifying parts of idiomatic expressions (UMR guidelines: all parts should be concatenated and used as 1 concept)

**PDT relations transformed to UMR concepts:**

Some phenomena captured as relations (edges) in PDT transformed to UMR **using new concepts (nodes)**:

* !! new concepts to cover **specific entities** **??? hotovo**
  + **contra** entity (CONTRA, as *Definitivní výsledek přišel v případu Hymowitz versus Lilly …* 'The definitive result came in the case of Hymowitz v.CONTRA Lilly')
  + **foreign-phrase** entity (FPHR)
  + **math** entity (OPER, intervals, etc.)

**Problem:** Structured data represented in UMR as special "entities" (e.g., date-entity, further structured with attributes like day, month, year, century, etc) or "quantities" (e.g., monetary-quantity or temporal-quantity-quantity, both with the attributes quant and unit) mainly not identified in PDT yet.

**II.4 Identification of events**

**Verb predicates:**

* **all verb predicates** (i.e., lexical verbs, excluding modal and temporal auxiliaries) are treated as events, disregarding their "packaging" (as there are no clear (formal) criterion for distinguishing, e.g., statives in Czech)
  + PropBank-like lexicon for Czech covers **43%** of verb predicates (Hajič et al, 2024), see below
  + the rest based on the PDT-Vallex lexicon 🡪 PropBank-like lexicon must be extended
* **semimodals** must be identified
* **phase verbs** must be identified

(e.g., UMR: inchoative, completive, and continuative verbs) – NEVER as separate event,

only inform the aspect value

* **LVC**
* **?? stative verbs in reference and modifications as non-events ??**

**Non-verbal predicates:** … not transformed yet

* **eventive nouns** … derived from verbs / nouns with verbal counterparts
  + ?? -ní/-tí nouns (type *přijíždění*) … JŠ: Email from July 15, 2024 (without forms)
    - almost 30% without valency frames
    - almost 50% with a single valency frame
    - almost 25% with more frames
    - ?? A kdyby se zohlednily formy:
      * nom --> gen, poss, instr, od+2
      * acc --> gen, poss, instr, od+2
      * ostatní formy by měly zůstat beze změny, příp. může nějaká u substantiva chybět či naopak přebývat.
  + ?? nominal events (type *příjezd*; type *volby, analýza*; ???)
  + ?? agentive nouns (type *učitel, volič*) (cs: činitelská) -> inverse roles
* **eventive adjectives**
  + - ?? type *(byl) unavený* (type *unaven* as passive participle, thus verb (MorfFlex))
    - ?? type *přijíždějící*
* **eventive adverbs**

Sources: MorfFlex, DeriNet (a data od Hanky), PDT-Vallex, SynSemClass (Eva Fučíková)

**Abstract rolesets:** … not identified yet

* **abstract predicates/rolesets**:
  + *být* /
  + *mít* /
  + patřit ‘belong’:
    - patřit-001 (v-w3411f6\_ZU, which substitutes v-w3411f2, v-w3411f5\_ZU ... náležet, přináležet, příslušet, být ve vlastnictví)

--> belong-91 ... ACT (possessum) --> ARG1, PAT (possessor) --> ARG2

* + - patřit-002 (v-w3411f3) ... frazem, ponechat (To ti patří!)
    - patřit-003 (v-w3411f1 ... náležet, řadit se, přináležet, být součást, spadat)

--> include-91 ... ACT (subset) --> ARG1, DIR3 (superset) --> ARG2

* + - patřit-004 (v-w3411f4 ... dát, umístit)

--> have-place-91 ... ACT (entity) --> ARG1, DIR3 (location) --> ARG2

* + - patřit-005 (v-w3411f7\_ZU) ... patří na+4 (asi význam zírat, nevidím v Teitoku), ponechat
  + vlastnit ‘own’:
    - vlastnit-001 (v-w7650f1, držet, spravovat)

--> have-91 ... ACT (possessor) --> ARG1, PAT (possessum) --> ARG2 etc.

* + ??? other verbs … should be converted to abstract predicates
  + other candidate construction should be identified like *Mariina/její taška*, ‘Maria’s/her bag’
* special **linguistic constructions** (e.g., have-degree-91, include-91) and
* **reifications**

**II. 5 Named Entities** … not available in PDT, cannot be transferred **(completely ignored so far!!)**

**Identification and classification of NEs:**

* UMR abstract concepts:
  + list: <https://docs.google.com/spreadsheets/d/1PVxgXW3ED3OWLieie9scr6iq_xuQ5RAA8YJKwbLwJ2E/edit?gid=0#gid=0>
  + definitions:

<https://docs.google.com/document/d/1Wx2jXRTosH3I8aDhdrxqYRH8TPABD3m1HuYSXivdAAg/edit?tab=t.0>

* PDT-C:
  + ?? names of persons
  + ?? other types
* **NameTag 3 Model** 
  + based on [Czech Named Entity Corpus 2.0 (Ševčíková et al., 2007)](https://ufal.mff.cuni.cz/cnec/cnec2.0).
  + <https://ufal.mff.cuni.cz/nametag/3/models#czech-cnec2>
  + **too coarse-grained** (compared to UMR)

**NEs anchoring: ???**

**II.6 UMR attributes**

**1. Aspect:** Can/Should be transferred. **??? hotovo**

**2. Polarity:**

* All types of flags/markers indicating negation are collected in the polarity attribute of the relevant concept (value "–").
* PDT grammateme negation = neg1

as in *nezralost dítěte* 'immaturity of a child ' [lemma=zralost 'maturity', negation=neg1]

* PDT grammateme indeftype = negat

as in negative pronouns/pronominal adverbs *nikdo* 'no one', *nikde* 'nowhere'

* PDT syntactic negation (negation morpheme *ne-* or negation particles *ne/nikoli(v)*)
* PDT negative interjection clauses (*Ne, ještě nepřišel.* 'No, he has not come yet.')
* **questions** (umr-unknown, truth-value) … not processed yet
* **embedded interrogative clauses** (truth-value) … not processed yet

**3. Mode:** Can/Should be transferred. **??? hotovo**

**4. Polite:** Can/Should be transferred for a portion of the data **??? hotovo**

* cannot be simply detected for the rest

**5. Refer:** Can/Should be transferred for a portion of the data

* based on morphological form for the rest ???

**POZOR** … číslo se propisuje k mnoha jednotkám, kde nemá být, jako adj, číslovka apod.

… ??? ponechat jen u **sempos ~ 'n.\*'** ???

… ??? kde nejsou gramatémy … morf. tag **~ 'P[PH5DZLWKQ].\*** ???

(např. cmpr9410\_001.umr, věta 3: *upravující, vypověditelný, všechen*)

**POZOR** … osoba se propisuje k mnoha jednotkám, kde nemá být, jako je např. sloveso

… ??? ponechat jen u **sempos ~ 'n.pron.** **def.pers.\*|n.pron.indef'** ???

cmpr9410\_001.umr, věta 4: *zakotvovat, mít*

**6. Degree** … not transformed yet

* list of intensifiers, downtoners, equals not available for Czech!!

**7. Quant** … not transformed yet

* how such structures can be identified in PDT-C ???

**8.** **Modal-strength** … not transformed yet

??? complicated interplay between:

* + sentmod (enunc VS. excl, desid, imper, inter), factmod grammateme MOD
  + factmod (asserted VS. appeal, potential VS. irreal) … jen PDT
  + deontmod (decl VS. muset, smět, chtít) … mimo WSJ
  + negation
  + semimodals (zakázat)
  + MOD expressions (asi, možná, zřejmě, pravděpodobně … cca 90 různých)

**Scope for quantification and negation** … not annotated**… sect. 3.1.5**

* NOT found in Eng. UMR 1.0

**III. Document Level Representation**

**III.1 Coreference**

* entity coreference
  + intra-sentence relations … done, see above
  + inter-sentence coreference … done
  + ??? bridging anaphora
* identify coreferential relations … only sporadically available in PDT

**III.2. Temporal relations** … not transformed yet

**III.3 Modality** … not transformed yet