**Czech UMR data, version 0.5**

**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)
* **apposition** … as identity-91
* **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*

* + - 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!!
  + inverse roles
    - relative clauses

**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

**II.3 Identification of events**

* **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
* **eventive nouns, adjectives, adverbs** … not transformed yet
* **abstract rolesets** … not identified yet

**II.4 Relations labeling**

**Verb-specific arguments labeling:**

* verb specific conversion for **43%** of verb predicates (= frames ≈ rolesets with arguments) (Hajič et al, 2024)
* the rest based on the PDT-Vallex lexicon

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

* type of NE (see the list [here](https://docs.google.com/spreadsheets/d/1PVxgXW3ED3OWLieie9scr6iq_xuQ5RAA8YJKwbLwJ2E/edit?gid=0#gid=0) and definitions [here](https://docs.google.com/document/d/1Wx2jXRTosH3I8aDhdrxqYRH8TPABD3m1HuYSXivdAAg/edit?tab=t.0))
* NE 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 "–").
* questions (umr-unknown, truth-value) and 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 **??? hotovo**

* based on morphological form for the rest ???

**6. Degree** … not transformed yet

**7. Quant** … not transformed yet

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

**Scope for quantification and negation** … not annotated

**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

**=====================================================================================**

**TO BE DONE / IMPERFECT CONVERSION / MANUAL ANNOTATION NEEDED**

**II. Sentence Level Representation**

**II.1 Structural changes**

* coordination … POZOR, někde víc ARG než 2 (cca 30) případů
  + "but-91", ??? kde ještě ???

**II.2 Nodes labeling**

* **supertypes**
* **POZOR: nejasné, co s qcomplex #Ast, #Period3, #Comma, #Colon, #Dash**

**II.3 Identification of events**

**Verb predicates:**

* the **Czech 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 ??**

**Eventive nouns, adjectives, adverbs … not transformed yet**

* **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
* adjectives derived from verbs
  + - ?? type *(byl) unavený* (type *unaven* as passive participle, thus verb (MorfFlex))
    - ?? type *přijíždějící*
* ?? adverbs derived from verbs
* Sources: MorfFlex, DeriNet (a data od Hanky), PDT-Vallex, SynSemClass (Eva Fučíková)

**Abstract predicates … not transformed 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.4 Relations labeling**

**Verb-specific arguments labeling:**

* verb-specific conversion must be extended
* disambiguation needed(ca 25 frames with two possible mappings, mail JŠ, 9 May, 2024)

**Default conversion table:**

* translation of some roles still too coarse (refinement needed)

**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. 5 Named Entities … not available in PDT, cannot be transferred**

* identification of NEs … **NameTag 3 Models**:

<https://ufal.mff.cuni.cz/nametag/3/models#czech-cnec2>

* PDT-C:
  + ?? names of persons
  + ?? other types
* type of NE (see the list [here](https://docs.google.com/spreadsheets/d/1PVxgXW3ED3OWLieie9scr6iq_xuQ5RAA8YJKwbLwJ2E/edit?gid=1255922856#gid=1255922856) and definitions [here](https://docs.google.com/document/d/1Wx2jXRTosH3I8aDhdrxqYRH8TPABD3m1HuYSXivdAAg/edit?tab=t.0))
* NE anchoring

**II.6 UMR attributes**

**1. Aspect … OK ??? hotovo**

**2. Polarity (sentence level) … ???**

All types of flags/markers indicating negation are collected in the polarity attribute of the relevant concept:

* 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 … OK ??? hotovo**

**4. Polite … OK ??? hotovo**

**5. Ref … OK ??? hotovo**

* based on morphological form for the rest ???

**6. Degree** … not transformed yet … intensifier, downtoner, equal

???

**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