**UMR 2.0 - Czech UMR data: Release notes**

Jan Štěpánek, Markéta Lopatková\*, Daniel Zeman , Eva Fučíková, Michal Novák, Šárka Zikánová, Jan Hajič

\*lopatkova@ufal.mff.cuni.cz

The Czech UMRs have been automatically converted from the Prague Dependency Treebank – Consolidated ([PDT-C](https://ufal.mff.cuni.cz/pdt-c)), version 1.9. (December 20, 2024).

The Czech UMRs can be can be downloaded as a part of the UMR 2.0 data from the LINDAT-Clarin repository (see the Licence). In total, the data contain 175 429 Czech sentences with partial UMR annotation.

The conversion covers selected phenomena pertaining to the sentence level annotation (esp. structure of the graph, nodes and relations labeling, and PropBank-like argument structure for verbs). Further, coreference relations are identified, both intra- and inter-sentential ones.

**1. Basic Characteristics of the PDT-C data format and the UMR data format**

Basic characteristics of the two approaches and their brief comparison are introduced by Lopatková et al. (2024, Sect. 2). Basic ideas of the conversion are also sketched there, with stress on the identification of

* *phenomena that can be more-or-less directly transferred from the available Czech annotation to the UMR structures,*
* *phenomena that require specific treatment and detailed analysis but still can be transferred, and*
* *phenomena that are unavailable in PDT-C and thus necessitate new annotations.*

**2. Adjusting sentence graph structure: Coreference processing**

Different representation of different types of the coreference relation in PDT-C and UMR frameworks affects significantly the overall structure of sentence graphs (Lopatková et al, 2024, Sect. 2 and Sect. 3.3). In PDT, all coreferential expressions are typically represented by separate nodes that are interlinked by a special coreferential relation. Special attributes of this relation identify the type of the coreference (grammatical coreference / text coreference / relations of bridging anaphora) and the type of reference (e.g., specific or generic). On the other hand, UMR introduces several ways how to treat phenomena represented by coreferential chains in PDT.

**a. Re-entrance** (within a sentence). This representation is primarily used to cover the coreference relation within a sentence, especially:

* ***Anaphor is a personal or possessive pronoun*** (incl. reflexives): Separate PDT-C nodes for an anaphor and its antecedent within a sentence are merged during the transfer (type *Maria – she – the girl – her*; but also with a nominal anaphor as *Maria – the girl*).
* ***Arguments of raising and control verbs***, represented as separate nodes in PDT-C, are merged during the transfer as well (incl. cases without overtly expressed anaphor, type *Martin viděl Petra přicházet = Martin viděl Petra, jak Petr přichází.* 'Martin saw Peter coming = Martin saw Peter as Peter is coming.').
* ***Other******coreferring participants or non-participants*** within a sentence, represented as separate nodes in PDT-C, are also merged during the transfer (???).

In relatively rare cases, separate nodes are kept in UMRs, esp. if an anaphor is further modified.

**b. Inverse roles** (within a sentence). A relative clause typically contains a relative expression (like a pronoun, as *kdo* 'who', *co* 'what', *jaký, který* 'that, which' or a relative adverb, as *kde* 'where', *kdy* 'when', *jak* 'how'), which is in PDT-C marked as coreferential with the modified expression (in the governing clause). The relative expression typically serves as an argument or adjunct of the predicate of the relative clause. When transferring to UMR, the relative is merged with its antecedent (i.e., modified expression); the original parent node of the relative is attached to the antecedent with the relation inverse to the original one.

Compare the PDT-C representation (left) and UMR representation (right) of the following example:

*Student, který hraje na housle, ...* 'The student, who is playing the violin, …'. The relative *který* 'who' (serving as ARG0 of the verb *hrát* 'play') is merged with its antecedent *student* 'student'; the relative clause is attached to the modified concept *student* 'student' with the inverse (i.e., ARG0-of) relation.

(s / student 'student' (s / student 'student'

:mod (p / hrát-001 'play' :ARG0-of (p / hrát-001 'play'

:ARG0 (w / kdo) 'who' :ARG1 (v / housle 'violin'))

:ARG1 (v / housle 'violin')))

coref (s :same-entity w)

**c. Separate coreferring concept nodes** (within a sentence / crossing sentence boundary). All nodes with a coreferential link in PDT-C are collected and the respective pairs of coreferring nodes are added to the document-level annotation. The relevant relation between individual pairs is identified, reflecting (1) whether they refer to the same entity or to the same event and (2) whether their mutual relation is a relation of the identity (both nodes represent the same referent) or it is a relation between a set and its (proper) subset/event and its subevent.

In particular, separate nodes are kept in UMRs if an anaphor is further modified.

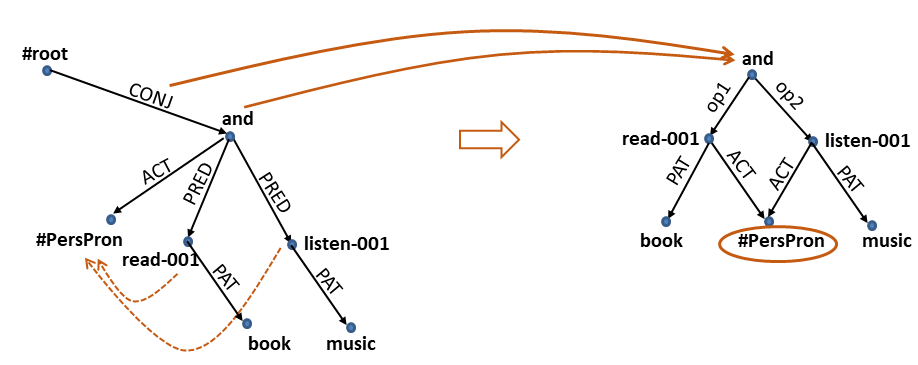
See Lopatková et al. (2024) for more examples.

**Warnings**:

* UMR does not recognize cataphoric reference (i.e., reference to an expression that appears later in the text). For all such instances, the relation is changed to the anaphoric one (and left for further analysis).
* Coreferential relations between events are only sporadically captured in the PDT-C thus this type is rare (and under-annotated) in the current version of the UMR data.
* The coreference relation between coordinated structures (see below) needs more detailed analysis. For this version, the UMR annotation reproduces the PDT-C annotation, i.e., the coreferential links interconnect the whole coordinated structures (not individual coordinated members).This representation should be refined in the future versions.
* #Rcp … coreference identified at the document level

**3. Adjusting sentence graph structure: Coordination and discourse relations, apposition**

**a. Coordination and discourse relations.** In general, representation of paratactic structures (coordination, discourse relations) follows the same principles in PDT-C and UMR: there is a special node in the graph for the whole paratactic structure (assigned with a discourse relation in UMR). Thus the transfer is more-or-less straightforward, dealing mainly with technicalities. Two points are relevant here:

* Participant and non-participant roles that are shared by coordinated concepts in PDT-C are represented as re-entrancies in UMR structures. See, e.g., the PDT-C (left) and UMR (right) graphs of the sentence *I read a book and listened to music* (both graphs are simplified) and the representation of the personal pronoun standing for "I".
* PDT-C, following Czech linguistic tradition, applies formal language-specific criteria for distinguishing coordinated and subordinated structures (e.g., both Czech conjunctions *protože, neboť* 'because' express causative relation, the former being classified as subordinating, the latter as coordinating); thus it is endowed with different roles for such structures (e.g., CAUS for the cause in subordinated relation and REAS for the same relation in coordinated structures).

As UMR should abstract from the formal means (applicable in a given language), the reification in 2 subtypes???

* ~~??? co je ta třetí poznámka??? … AHA, GRAD … tu ale až k nodes labeling~~
* 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)?

**Warning:**

* The coreference relation between coordinated structures needs more detailed analysis, see above.

=====================================================================================

**4. Nodes labeling**

??? explicitní rozlišování, co je event a co ne

(concepts = entities + events ; discourse relations as NON-events)

**5. Relations labeling**

**6. PropBank-like argument structure for verbs**

**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 constructions where both types can appear (e.g., constructions with the meaning of comparison, where 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