**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**

**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 below (II.3)

**II.3 Identification of events**

**Events expressed as "full" predicates:**

* **all verb predicates** are treated in the same way 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 predicates, reifications not transformed 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),
  + *být,* *mít* excluded in this phase
  + disambiguation needed … ca 25 frames with two possible mappings (mail JŠ, 9 May, 2024)

**Default conversion table:** (still some problems which need a refinement)

* several **new labels** to cover PDT-specific annotation:
  + !! **new** **roles**
    - **effect** (EFF)
    - **comparison** (based on CPR; complex structures, should be further inspected and refined)
    - **regard** (CRIT, REG)
    - **result** (RESL)
  + !! **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)

**PDT relations transformed to UMR concepts:**

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

* !! new concepts for MWE annotation
  + **predicative-noun** (CPHR; should be refined)
  + **part-of-phraseme** (DPHR)
* !! new concepts to cover specific entities
  + **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.)

**II. 5 Named Entities:**

* identification of NEs:
  + ?? names of persons
  + ?? other types
* ?? NEs anchoring

**II.6 UMR attributes**

**Aspect:**

verb predicates: diat grammateme (result)??

* a (tentative) list of verbs (= verb senses) expressing states

-> :aspect **state**

* verb usages with an iterative annotation (grammateme iter / note in MorfFlex)

-> :aspect **habitual**

* morphologically imperfective verbs / grammateme aspect = proc (processual)

-> :aspect **activity**

* morphologically perfective verbs / grammateme aspect = cpl (complex)

-> :aspect **performance**

* morphologically biaspectual verbs / / grammateme aspect = nr

-> :aspect: **proces**

* -91 rolesets:

-> :aspect **state**

**Polarity:**

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.')

**Ref**

* **refer-number**
* **refer-person**

**Mode:**

UMR:

* **imperative** (incl. exclamatory)
* **interrogative**
* **expressive** … for exclamational words such as *ah, ha, hmm, oh, wow, yippee* that express emotions, but don't refer to a clear event, object or property.

(not used for disfluency markers)

**Degree** … not transformed yet

**Quant** … not transformed yet

**III. Document Level Representation**

**1. Coreference** … see above for intra-sentence

* + ?? inter-sentence coreference
  + bridging anaphora
  + ?? identify coreferential relations among events

2. ?? **temporal** relations

3. ?? **modality** … modal-strength based on the deontmod grammateme

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

**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", ???

**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 lexicon must be extended
* semimodals, phase verbs

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

* **nouns** derived from verbs / nouns with verbal counterparts
  + - ?? -ní/-tí nouns (type *přijíždění*)
    - ?? 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

**Abstract predicates** … not transformed yet

* abstract predicates/rolesets:
  + být / mít / other verbs … should be converted to abstract predicates
  + other candidate construction should be identified

**II.4 Relations labeling**

Arguments:

* verb-specific