

From the Prague Dependency Treebank to the Uniform Meaning Representation: Gold-Standard Czech UMR Data and Partial Automatic Conversion

Markéta Lopatková, Hana Hledíková, Jan Štěpánek, Daniel Zeman

ITAT 2025, WAFNL, Sept. 2029, Telgárt, Slovakia

Motivation

- goal: Uniform Meaning Representation for Czech
 - semantics, abstracting away from syntax
 - cross-linguistic applicability
 - broad sem. interpretation of the text for cross-lingual applications
- annotation from scratch:
 - time consuming
 - expertise and training
- re-use existing corpus:
 - **automatic conversion** from Prague Dependency Treebank
 - rich annotation already there
 - the same procedure for all languages with PDT annotation
 - expertise and training still needed
 - evaluation: **manually annotated data**

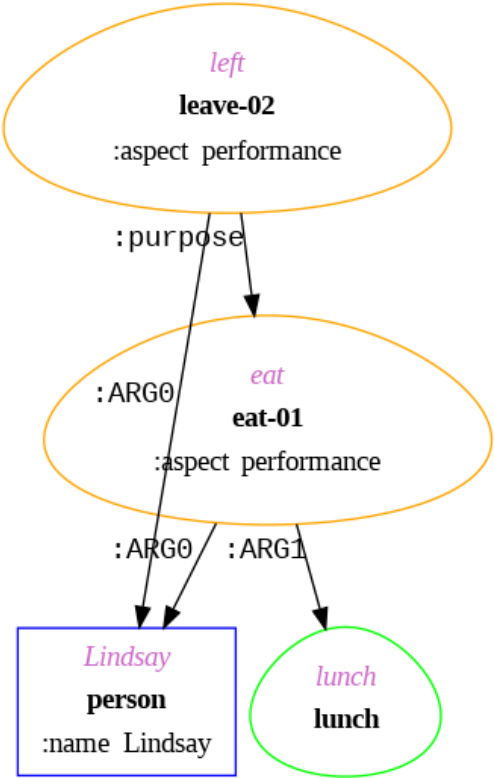
Uniform Meaning Representation

#####

meta-info :: sent_id = u_tree-cs-s1-root
:: snt1
Index: 1 2 3 4 5 6 7 8
Words: Lindsay left in order to eat lunch .

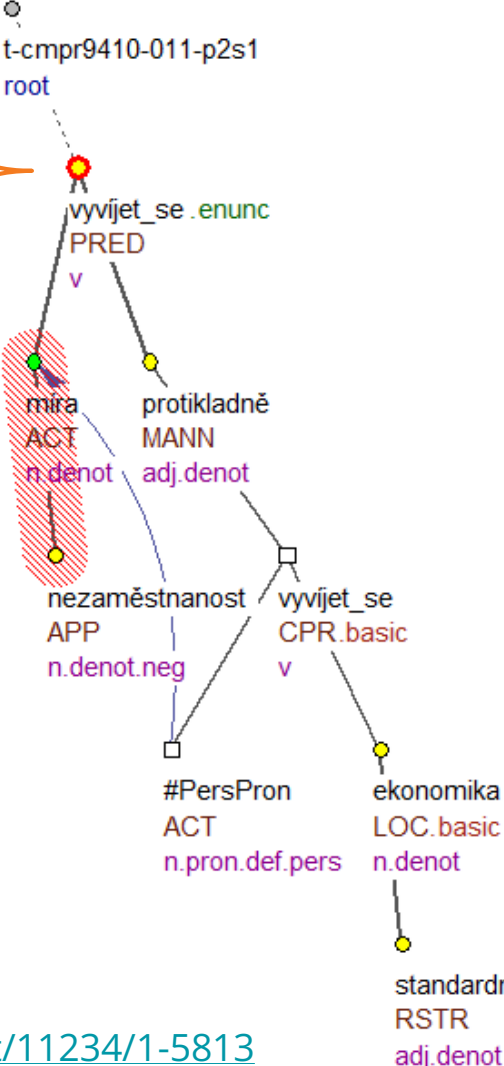
sentence level graph:
(s1l / leave-02
 :ARG0 (s1p / person
 :name (s1n / name :op1 "Lindsay"))
 :aspect performance
 :purpose (s1e / eat-01
 :ARG0 s1p
 :ARG1 (s1l2 / lunch)
 :aspect performance))
alignment:
s1l: 2-2 s1p: 1-1 s1n: 0-0 s1e: 6-6 s1l2: 7-7

document level annotation:
(s1s0 / sentence
 :temporal ((document-creation-time :before s1l)
 (s1l :after s1e))
 :modal ((root :modal author)
 (author :full-affirmative s1l)
 (author :full-negative s1e)))



Prague Dependency Treebank

a	Structure
aux.rf	Unordered list
	a#a-cmpr9410-011-p2s1w3
	a#a-cmpr9410-011-p2s1w5
	a#a-cmpr9410-011-p2s1w4
lex.rf	a#a-cmpr9410-011-p2s1w6
deepord	4
functor	PRED
gram	Structure
aspect	proc
deontmod	hrt
diatgram	act
factmod	potential
iterativeness	it0
sempos	v
tense	nil
id	t-cmpr9410-011-p2s1w6
nodetype	complex
sentmod	enunc
t_lemma	vyvíjet_se
tfa	f
val_frame.rf	v#v41pgqA



Míra nezaměstnanosti by se měla vyvíjet protikladně než ve standardní ekonomice.

The unemployment rate should develop in the opposite direction to that in a standard economy.

UMR vs. PDT

UMR

- represents meaning

PDT

- represents linguistically structured meaning (vs. situational meaning)
- topic-focus articulation

UMR vs. PDT

UMR

- represents meaning
- encodes the frame-based predicate-argument structure of all eventive concepts

PDT

- represents linguistically structured meaning (vs. situational meaning)
- topic-focus articulation
- predicate-argument structure (valency) and dependency relations

UMR vs. PDT

UMR

- represents meaning
- encodes the frame-based predicate-argument structure of all eventive concepts
- for each event, complex information
 - aspect
 - temporal chains
 - epistemic modality

PDT

- represents linguistically structured meaning (vs. situational meaning)
- topic-focus articulation
- predicate-argument structure (valency) and dependency relations
- meaning of individual morphological categories

UMR vs. PDT

UMR

- represents meaning
- encodes the frame-based predicate-argument structure of all eventive concepts
- for each event, complex information
 - aspect
 - temporal chains
 - epistemic modality
- coreference

PDT

- represents linguistically structured meaning (vs. situational meaning)
- topic-focus articulation
- predicate-argument structure (valency) and dependency relations
- meaning of individual morphological categories
- coreference

Towards Gold-Standard UMR Data for Czech

Towards Gold-Standard UMR Data for Czech

Gold-standard data:						
(sub)corpus	sentences	tokens	tokens per sentence	PDT-C nodes	UMR nodes	UMR per PDT-C
PDT	25	467	18.7	378	375	0.99
PDTSC	50	374	7.5	321	442	1.38
PCEDT	16	474	29.6	400	307	0.77
total	91	1315	14.5	1099	1124	1.02



Towards Gold-Standard UMR Data for Czech

Gold-standard data:						
(sub)corpus	sentences	tokens	tokens per sentence	PDT-C nodes	UMR nodes	UMR per PDT-C
PDT	25	467	18.7	378	375	0.99
PDTSC	50	374	7.5	321	442	1.38
PCEDT	16	474	29.6	400	307	0.77
total	91	1315	14.5	1099	1124	1.02

Parallel annotations:						
(sub)corpus	sentences	tokens	tokens per sentence	PDT-C nodes	UMR nodes Annot1 / Annot2	UMR per PDT-C (avg.)
PDT	11	192	17.5	151	153 / 150	1.00
PDTSC	10	63	6.3	68	75 / 71	1.26
total	21	255	12.1	209	228 / 221	1.07



Inter-Annotator Agreement (IAA)

UMR graphs = as a set of triples (x, y, z) :

- (node, relation, node)
- (node, attribute, value)

Metric for graph comparison:

1) Match nodes:

- different number of nodes
- different alignment (nodes to words)

 *ju:mætʃ*

- maps nodes primarily by word alignment
- for nodes without alignment, requires concept identity
- forces 1:1 mapping (selected the "best" node from 1:N)

2) Similarity is measured as the **F₁-score** of the triples

Towards Gold-Standard UMR Data for Czech

- final IAA (after reconciliation; table taken from Štěpánek et al., 2025)

UMR node mapping:					
Annot1 nodes	Annot2 nodes	mapped	recall	precision	F_1
228	221	215	94%	97%	96%
Concept and relation comparison (only mapped nodes):					
Annot1 triples	Annot2 triples	match	recall	precision	F_1
633	644	595	94%	92%	93%
Concept and relation comparison:					
Annot1 triples	Annot2 triples	match	recall	precision	$jumæff = F_1$
663	659	595	90%	90%	90%

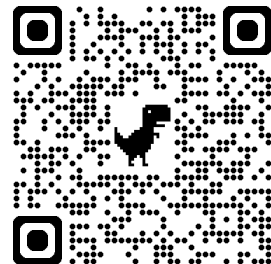
- analysis of main mismatches in the paper (events and argument structure, ellipses, granularity of NE classification, relations vs. attributes, attributes and their values)
- UMR allows for multiple valid annotations of the same meaning !!!

Automatic (Partial) Conversion

Automatic (Partial) Conversion

based on the tectogrammatical structure:

- structural transformations
 - coordination
 - coreference
 - relative clauses
 - raising and control verbs
- node labels
 - t_lemma → concept
- edge labeling
 - valency lexicon → PropBank
default table
- selected attributes
 - aspect
 - refer-person, refer-number
 - degree, polarity, quant
- node alignment



Automatic (Partial) Conversion

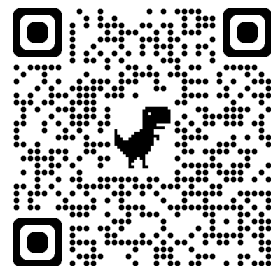
based on the tectogrammatical structure:

- structural transformations
 - coordination
 - coreference
 - relative clauses
 - raising and control verbs
- node labels
 - `t_lemma` → `concept`
- edge labeling
 - valency lexicon → PropBank default table
- selected attributes
 - aspect
 - refer-person, refer-number
 - `degree`, `polarity`, `quant`
- node alignment

often interact




further increases
the conversion complexity



Automatic (Partial) Conversion

based on the tectogrammatical structure:

- structural transformations
 - coordination
 - coreference
 - relative clauses
 - raising and control verbs
- node labels
 - `t_lemma` → `concept`
- edge labeling
 - valency lexicon → PropBank default table
- selected attributes
 - aspect
 - refer-person, refer-number
 - `degree`, `polarity`, `quant`
- node alignment

often interact  further increases the conversion complexity

ignored (so far)

- attributes:
 - `mode`, `polite`
 - quote, modal-strength
 - `wiki`
- most of the document level annotation
 - temporal
 - modal

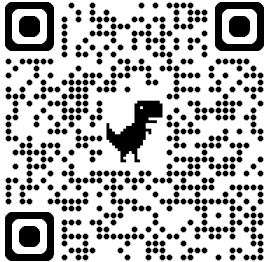
Automatic (Partial) Conversion – Quantitative Comparison

- Automatic conversion: (tables taken from Štěpánek et al., 2025)

UMR node mapping:						
corpus	MAN nodes	AUTO nodes	mapped	recall	precision	F ₁
PDT	375	349	284	76%	81%	78%
PDTSC	442	305	235	53%	77%	63%
PCEDT	307	327	244	79%	75%	77%
total	1124	981	763	68%	78%	72%

- Manual annotation (inter-annotator agreement):

UMR node mapping:					
Annot1 nodes	Annot2 nodes	mapped	recall	precision	F ₁
228	221	215	94%	97%	96%



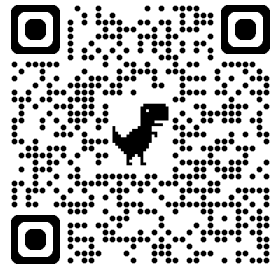
Automatic (Partial) Conversion – Quantitative Comparison

- Automatic conversion: (tables taken from Štěpánek et al., 2025)

Concept and relation comparison (only mapped nodes):						
corpus	MAN triples	AUTO triples	match	recall	precision	F ₁
PDT	844	819	502	59%	61%	60%
PDTSC	622	633	352	57%	56%	56%
PCEDT	714	588	342	48%	58%	53%
total	2180	2040	1196	55%	59%	57%

- Manual annotation (inter-annotator agreement):

Concept and relation comparison (only mapped nodes):					
Annot1 triples	Annot2 triples	match	recall	precision	F ₁
633	644	595	94%	92%	93%



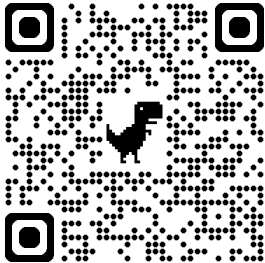
Automatic (Partial) Conversion – Quantitative Comparison

- Automatic conversion: (tables taken from Štěpánek et al., 2025)

Concept and relation comparison:						
corpus	MAN triples	AUTO triples	match	recall	precision	$ju.mæff = F_1$
PDT	1082	916	502	46%	55%	50%
PDTSC	1318	770	352	27%	46%	34%
PCEDT	916	757	342	37%	45%	41%
total	3316	2443	1196	36%	49%	42%

- Manual annotation (inter-annotator agreement):

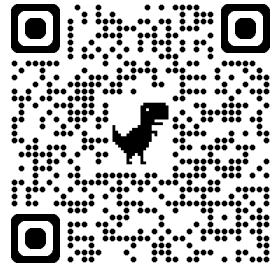
Concept and relation comparison:						
Annot1 triples	Annot2 triples	match	recall	precision	$ju.mæff = F_1$	
663	659	595	90%	90%	90%	



From PDT to UMR:

Gold-Standard Czech UMR Data and Partial Automatic Conversion

- two different meaning representations
- manually annotated Czech UMR gold-standard data
 - IAA 90 %
- evaluation of the automatic (partial) conversion
 - transforms selected language phenomena from PDT to UMR
 - 53-60% accuracy on the aligned nodes
 - plan: cover more phenomena in the (near) future
- automatic conversion as an essential first step to reduce costs for full manual annotation



Acknowledgements



- The work described herein has been supported by the following grants:
 - Czech Science Foundation, *Language Understanding: from Syntax to Discourse* (Project No. 20-16819X)
 - Ministry of Education, Youth, and Sports of the Czech Republic, *LINDAT/CLARIAH-CZ* (Project No. LM2023062)
- The project has been using data and tools provided by the *LINDAT/CLARIAH-CZ Research Infrastructure* (<https://lindat.cz>), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).

