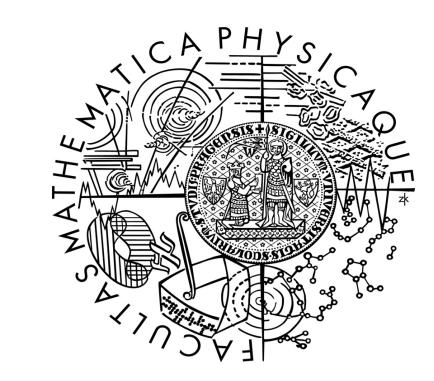


Mapping Czech Verbal Valency to PropBank Argument Labels





Jan Hajič, Eva Fučíková, Markéta Lopatková, Zdeňka Urešová

Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics (ÚFAL MFF UK)



Goal and Research Questions

We present here a high-coverage pilot study of mapping the labeling system used in PropBank (for English) to PDT-based annotation (for Czech), which has so far used mainly valency lexicons for annotation projects, under different levels of specification and different theoretical assumptions.

The purpose of this study is both theoretical (comparing the argument labeling schemes) and practical (to be able to annotate Czech under the standard UMR specifications).

Lexical Resources

PropBank Frame Files (Palmer et al., 2005) CzEngVallex bilingual valency lexicon (Urešová et al., 2016) (with monolingual PDT-Vallex and EngVallex)

SynSemClass ontology 5.0 (Urešová et al., 2023)

Paralel Corpus

Source Data

Czech English Dependency Treebank (PCEDT) (Hajič at al, 2012)

- PropBank annotation for EN
- PDT-based annotation for both CS and EN

The Mapping Algorithm

 ${f 1}$. For each Czech verb (sense), collect its valency frame from the valency lexicon (PDT-Vallex) & its semantic class and semantic roles (SynSemClass)

2. Collect its PCEDT translation equivalents (CzEngVallex) and all equivalents from the same class (SynSemClass) **3.** Collect available argument mappings (CzEngVallex, SynSemClass, EngVallex-PropBank mapping) and PCEDT corpus frequencies

4. Consolidate the information and create the new, complete **PropBank-style roleset** for the given Czech verb sense

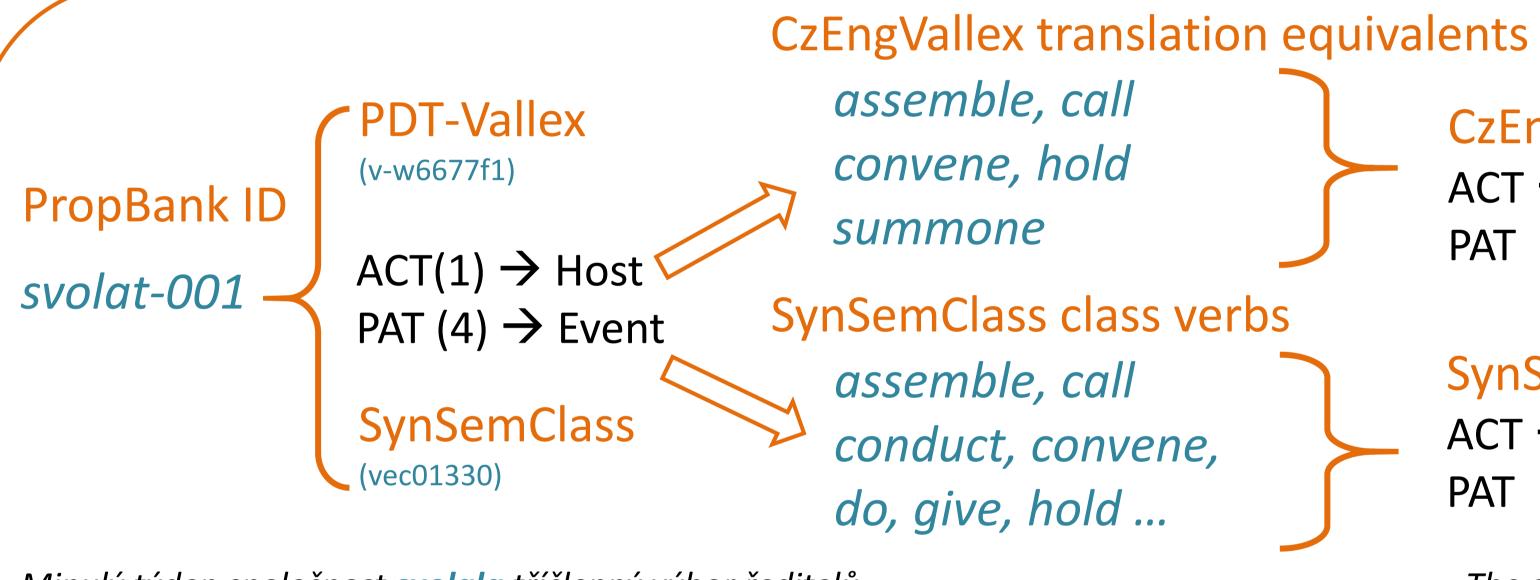
consolidated mapping

 $ACT \rightarrow ARG0 \dots Host$

 $PAT \rightarrow ARG1 \dots Event$

for svolat-001:

Example – Successful Mapping (verb svolat "assemble; call; convene")



Minulý týden společnost **svolala** tříčlenný výbor ředitelů,

Kvůli hlasování o návrzích svolala BNL na prosinec valnou hromadu akcionářů. Křesťanský vůdce Libanonu **svolal** krizovou schůzi kabinetu poté, co zjistil ... Thatcherová v Británii svolala hlavní poradce ke strategickým jednáním, ...

 $ACT \rightarrow Host \rightarrow ARGO$ (310 times), ARG1 (1 time) $PAT \rightarrow Event \rightarrow ARG1$ (418 times)

SynSemClass mapping

CzEngVallex mapping

 $ACT \rightarrow ARGO$ (141 times)

 $PAT \rightarrow ARG1$ (197 times)

The company last week **assembled** a three-member committee of directors ...

BNL called a shareholders' assembly meeting in December to vote on the proposals. Lebanon's Christian leader convened an emergency meeting of his cabinet after indications ... Britain's Thatcher **summoned** senior advisers for strategy talks ...

President Bush and Soviet leader Mikhail Gorbachev will hold an informal meeting in early December, ...

Examples – Unsuccessful Mappings

PDT-VALLEX (v-w90f1) asistovat-001

ACT(1) PAT(3)

SynSemClass

(vec00067)

CzEngVallex mappings assist

support $ACT \rightarrow ARGO$ (16 times) ARGO (102 times)

ARG1 (149 times) PAT \rightarrow ARG1 (20 times), ARG2 (3 times)

SynSemClass mapping

ACT \rightarrow Protagonist \rightarrow ARG0 (166 times), ARG1 (128 time), ARG2 (1 time) PAT \rightarrow Event \rightarrow ARG1 (53 times), ARG2 (295 times)

hnát-001 ACT(1) PAT(4)DIR3(=)

SynSemClass

(vec00821)

PDT-VALLEX

(v-w1067f2)

CzEngVallex mapping carry, drive, run up, take, underpin

 $ACT \rightarrow ARGO$ (57 times)

 $PAT \rightarrow ARG1$ (97 times) DIR3 → ---

discontinuous ARGS

SynSemClass mapping

ACT \rightarrow Stimulus \rightarrow ARG0 (164 times), ARG1 (1 time)

PAT \rightarrow Affected \rightarrow ARG1 (232 times)

DIR3 -> State_final -> ARG1 (19 times), ARG2 (22 times)

ARG3 (37 times)

Statistics: arguments (in fully mapped rolesets)

	unamb-	pref-	un-	
	iguous	erred	mapped	total
functors	9,465	8,579	24,072	42,116
percent	23	20	57	100

Statistics: rolesets

	auto-	un-	
	suggested	assigned	total
rolesets	5,085	10,569	15,654
percent	32	68	100

Outputs

ambiguous role

mapping

PCEDT

- PropBank style rolesets for more than 5,000 Czech verbs (= verb senses) obtained automatically as unambiguous or preferred mappings from different lexical resources.
- Collected data with the necessary valency / predicate-argument information and clickable links for annotators to finish the work manually in an efficient manner.
- The resulting Czech PropBank frame files for future Czech UMR annotation requiring PropBank-style argument labels.
- Possibility for more direct, large-scale comparison between the two approaches to predicate-argument labeling.

Supported by the grant Language Understanding: from Syntax to Discourse of the Czech Science Foundation (Project No. 20-16819X). We acknowledge the use of data provided the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), LREC – DMR 2024 supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).