Parallel Dependency Treebank Annotated with Interlinked Verbal Synonym Classes and Roles

Zdeňka Urešová, Eva Fučíková, Eva Hajičová, Jan Hajič
Charles University, Prague, Czech Republic

Outline

- Semantic annotation enrichment of PCEDT via CzEngClass
  - Automatic preprocessing + manual correction
- Resources used
  - PCEDT
  - CzEngClass
- Annotation
  - Semantic attributes assignment (class + roles)
  - Automatic pre-annotation
  - Disambiguation, Corrections and Analysis
- Conclusions and Future Work

(Lexical) Sematic Annotation

- PCEDT Parallel treebank: Czech/English (PTB/WSJ)
  - Richly annotated treebank (morphology, syntax, SRL, coreference)

+

- Bilingual, verb synonym lexicon CzEngClass

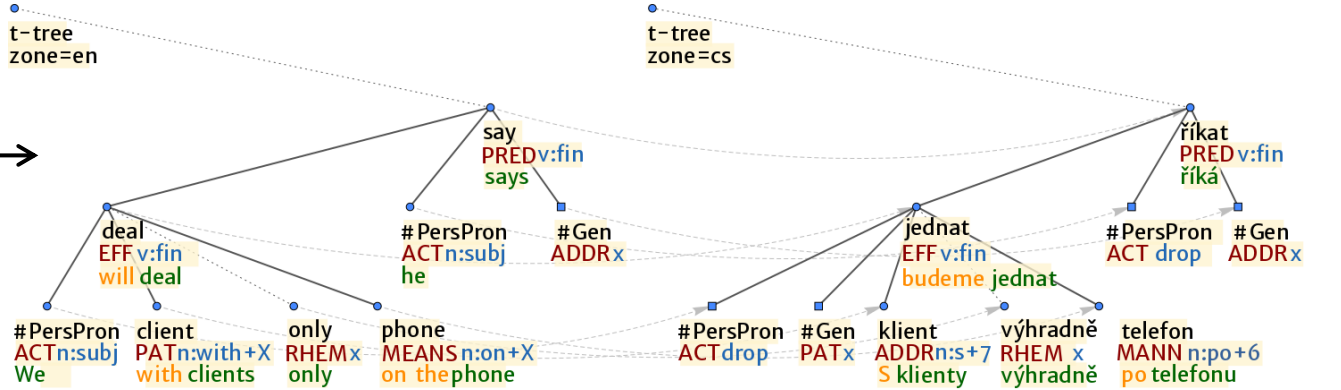
- Parallel treebank, annotated with reference links to CzEngClass and many other semantic lexicons
  - direct reference links from the verbs in the corpus
    - Granularity: to the individual lexicons' (sub)entries

The PCEDT treebank

- Prague Czech-English Dependency Treebank
  - https://catalog.ldc.upenn.edu/LDC2012T08
  - Searchable at
    - https://lindat.mff.cuni.cz/services/pmltq/#!/treebank/pcedt20_cz/query/
  - 55,000 sentences on each language side
    - Annotated: Tectogrammatical Representation (FGD)
      - Dependency-based, syntactic-semantic layer annotation
        - Also morphology, syntax
    - Content verbs sense- and valency-annotated
      - PDT-Vallex (Czech), EngVallex (English) lexicons

PCEDT and Valency

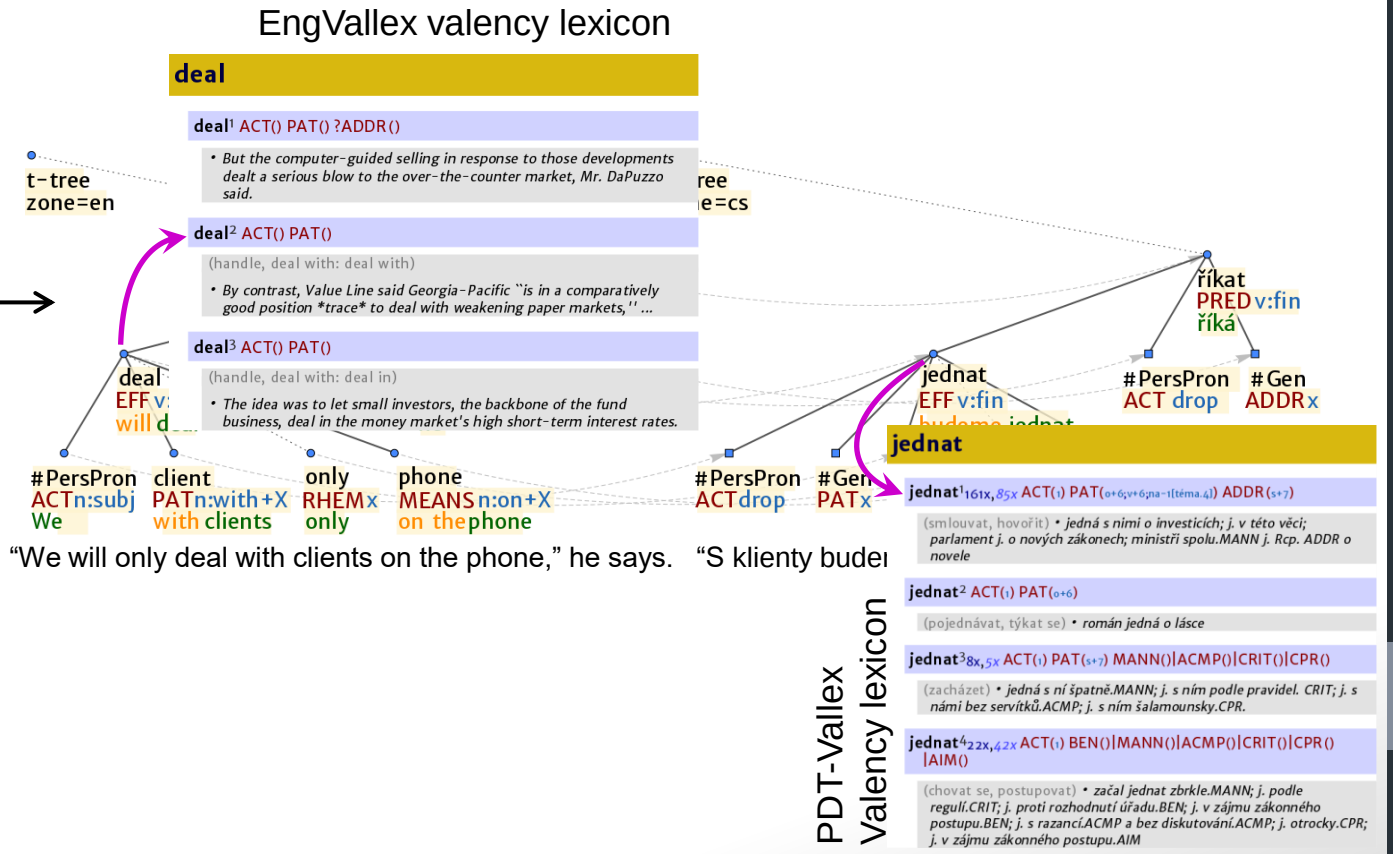
Tectogrammatical layer of annotation →



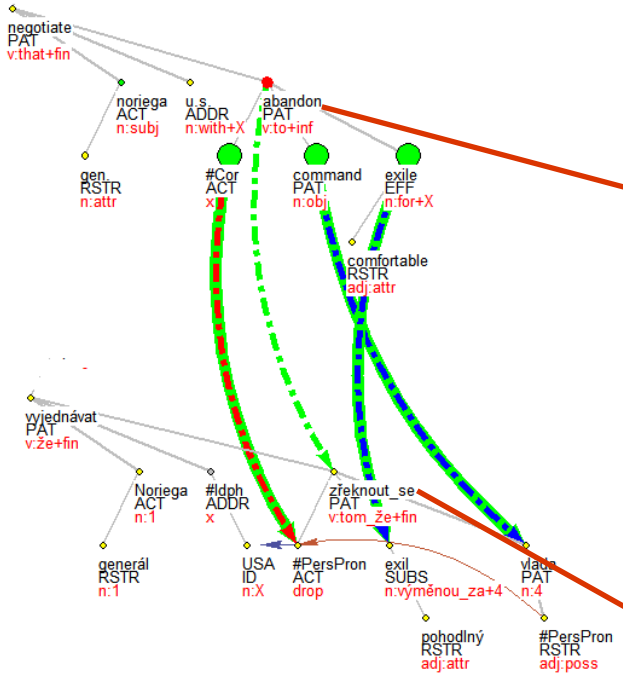
"We will only deal with clients on the phone," he says. "S klienty budeme jednat výhradně po telefonu," říká.

PCEDT and Valency

Tectogrammatical layer of annotation (manually created) →



Alignment: Verbs, Arguments



EN: ... Noriega […] to abandon his command for a comfortable exile
CS: ... Noriega […] že by se výměnou za pohodlný exil zřekl své vlády

EngVallex

abandon

abandon[1] **ACT**(sub) **PAT**(to[objpp,ving,a])

(reflexive)

- *Once he had abandoned himself to the very worst, once he had quieted all the dragons of worry and suspense, there would n't be very much for Mae to do.*

abandon[2] **ACT**() **PAT**() **EFF**()

- *One Colombian drug boss, upon hearing in 1987 that Gen. Noriega was negotiating with the U.S. [*] to abandon his command for a comfortable exile, sent him a hand-sized mahogany coffin engraved with his name.*

abandon[3] **ACT**() **PAT**()

(leave behind: typical transitive)

- *And they believe the Big Board, under Mr. Phelan, has abandoned their interest.*
- *John abandoned his pursuit of an Olympic gold medal as a waste of time.*

abandon[2]- zřeknout se
ACT →ACT
PAT→PAT
EFF →SUBS

PDT-Vallex

zřeknout se

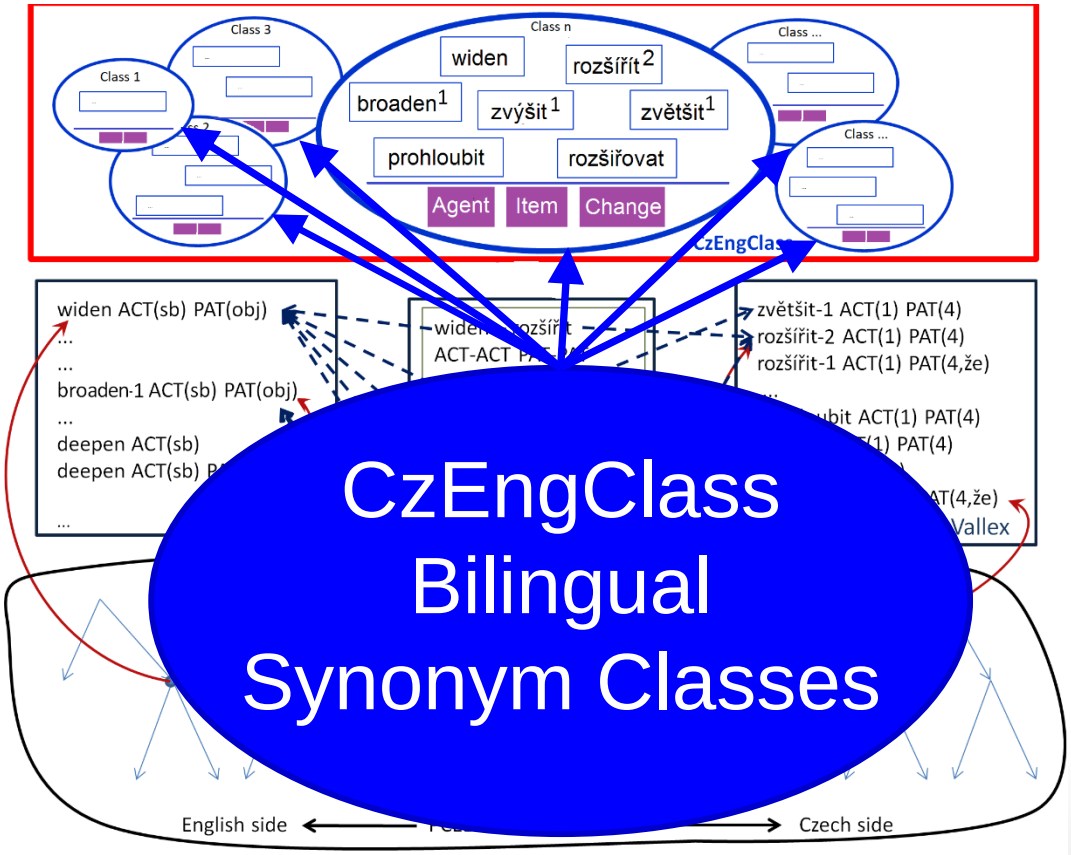
zřeknout se$_{7x,7x}$ **ACT**(1) **PAT**(2)

(zřící se, odmítnout, vzdát se) • *zřeknout se sestavení vlády*

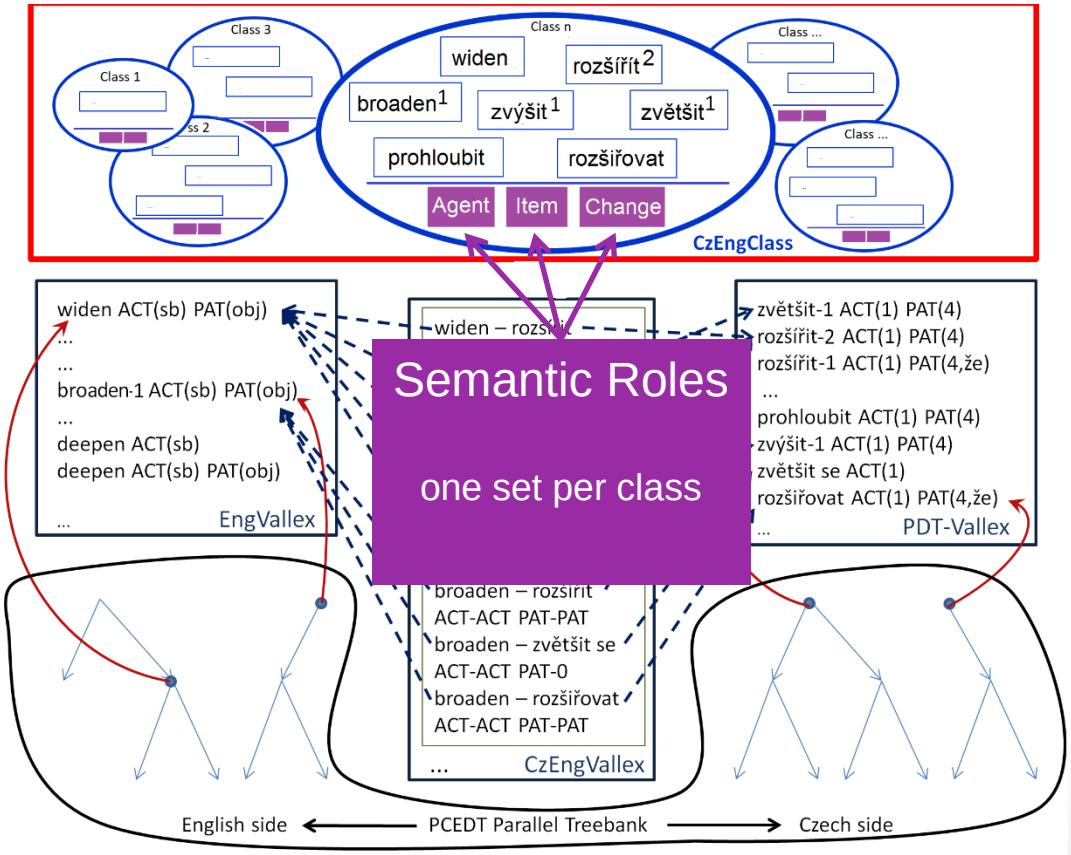
CzEngClass Contents

- Cross-lingual synonym classes (Cz/Eng, but…)
  - grouping verb senses with similar meaning - Class members
  - Common set of Semantic Roles per class (Roleset)
  - Mapping
    - valency arguments ↔ semantic roles, for each verb & argument
- Entries refer (link) to several existing semantic resources
  - Internal (keeps original valency frame IDs)
    - PDT-Vallex, EngVallex, and CzEngVallex
  - External
    - FrameNet, VerbNet, PropBank, OntoNotes, WordNet (Eng, Cz), Vallex
- SynEd – CzEngClass Lexicon Annotation editor
- Web version (upcoming – “beta” version for now, API)

CzEngClass: Structure

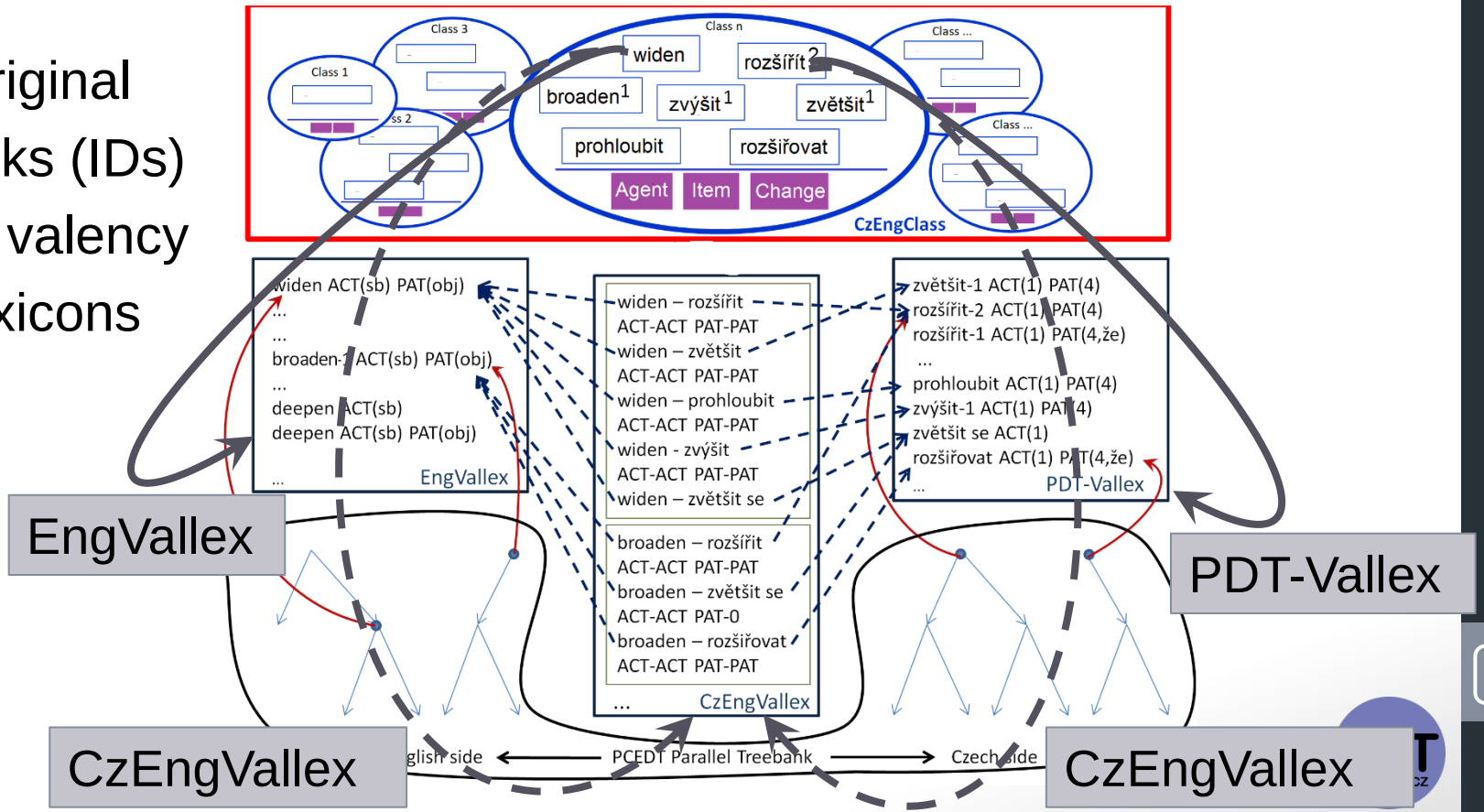


CzEngClass: Structure



CzEngClass: Structure

- Original links (IDs) to valency lexicons



CzEngClass: Structure

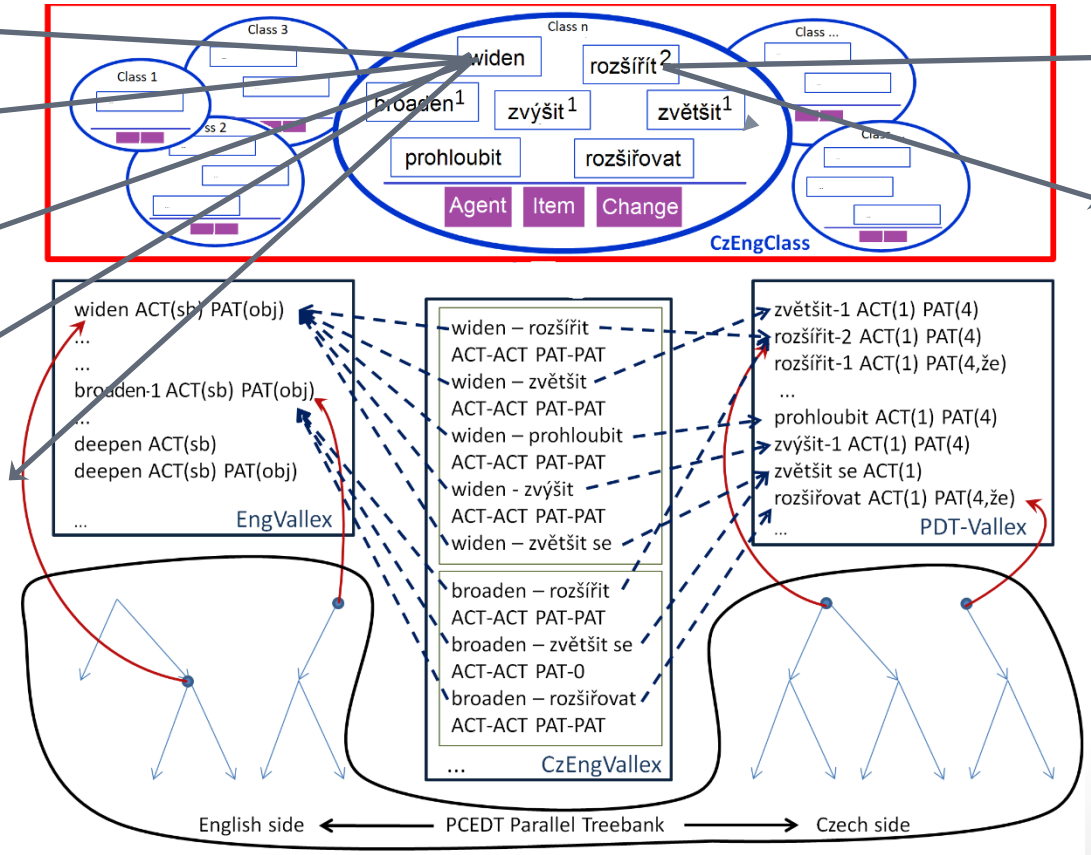
FrameNet

WordNet

VerbNet

PropBank

OntoNotes senses



Vallex

Czech WordNet

- Links to external lexicons

Example Synonym Class

stěžovat si (complain)	Complainer	Addressee	Complaint
complain	ACT	ADDR	PAT/EFF
gripe	ACT	ADDR	PAT
grumble	ACT	ADDR	PAT
brblat	ACT	LOC	PAT
postěžovat si	ACT	ADDR	PAT
protestovat	ACT	LOC	PAT
reptat	ACT	LOC	PAT
stěžovat si[1]	ACT	ADDR	PAT
stěžovat si[2]	ACT	ADDR	PAT/EFF

He.ACT complained to her.ADDR that her son lies. PAT
He.ACT complained to her.ADDR about her son.PAT that he lies.EFF

Synonym Class

Cz-Eng Synonym Class

Semantic Roles (Common Roleset)

stěžovat (complain)	Complainer	Addressee	Complaint
complain	ACT	ADDR	PAT/EFF
gripe	ACT	ADDR	PAT
grumble	ACT	ADDR	PAT
brblat	ACT	LOC	PAT
postěžovat si	ACT	ADDR	PAT
protestovat	ACT	LOC	PAT
reptat	ACT	LOC	PAT
stěžovat si[1]	ACT	ADDR	PAT
stěžovat si[2]	ACT	ADDR	PAT/EFF

He.ACT complained to her.ADDR that her son lies. PAT
He.ACT complained to her.ADDR about her son.PAT that he lies.EFF

Synonym Class

Cz-Eng Synonym Class

stěžovat (complain)	Complainer	Addressee	Complaint
complain	ACT	ADDR	PAT/EFF
gripe	ACT	ADDR	PAT
grumble	ACT	ADDR	PAT
brblat	ACT	LOC	PAT
postěžovat si	ACT	ADDR	PAT
protestovat	ACT	LOC	PAT
reptat	ACT	LOC	PAT
stěžovat si[1]	ACT	ADDR	PAT
stěžovat si[2]	ACT	ADDR	PAT/EFF

Semantic Roles (Common Roleset)

Class Member(s)

He.ACT complained to her.ADDR that her son lies. PAT
He.ACT complained to her.ADDR about her son.PAT that he lies.EFF

Synonym Class

Cz-Eng Synonym Class

stěžovat (complain)	Complainer	Addressee	Complaint
complain	ACT	ADDR	PAT/EFF
gripe	ACT	ADDR	PAT
grumble	ACT	ADDR	PAT
brblat	ACT	LOC	PAT
postěžovat si	ACT	ADDR	PAT
protestovat	ACT	LOC	PAT
reptat	ACT	LOC	PAT
stěžovat si[1]	ACT	ADDR	PAT
stěžovat si[2]	ACT	ADDR	PAT/EFF

Semantic Roles (Common Roleset)

Mapped Valency Frame(s)

Class Member(s)

He.ACT complained to her.ADDR that her son lies. PAT
He.ACT complained to her.ADDR about her son.PAT that he lies.EFF

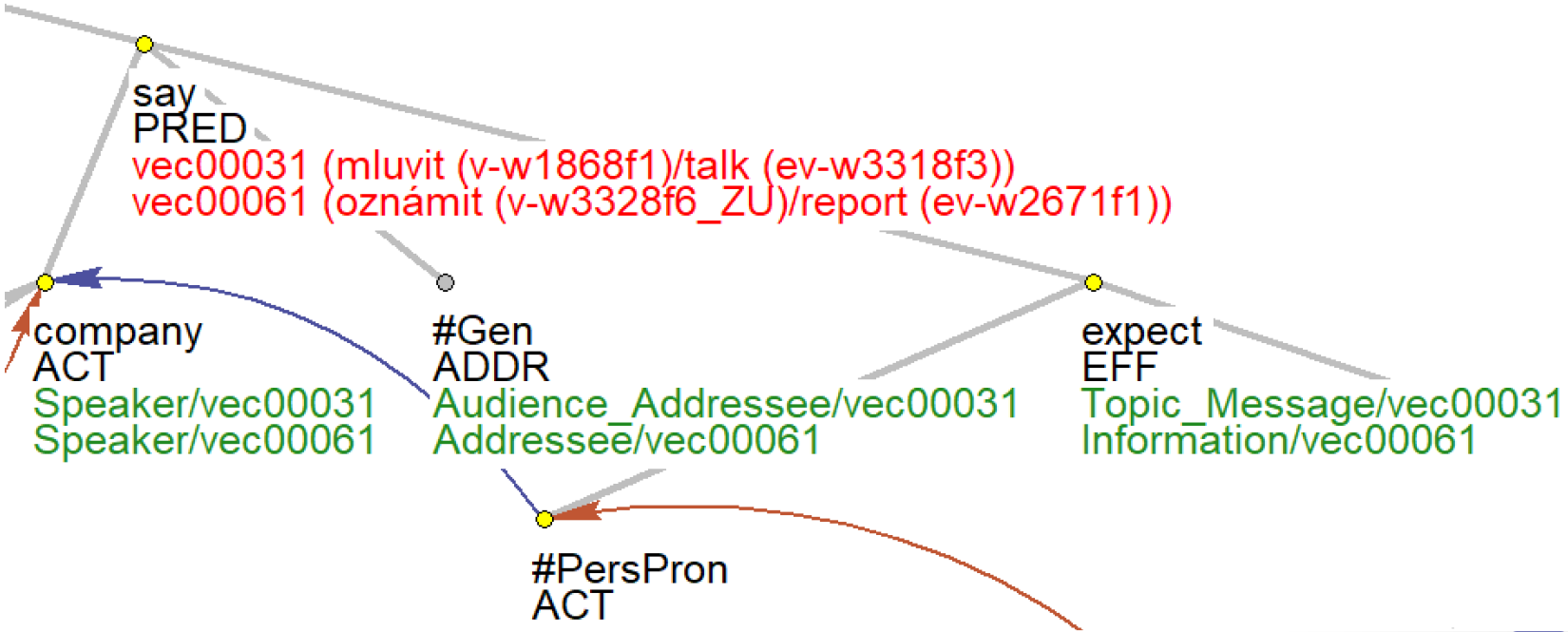
Annotation

- Goal: corpus with all events annotated by a high-coverage multilingual verbal synonym lexicon entries (= CzEngClass)
- Used for
  - Theoretical studies (lexical semantics, translatology, corpus annotation, etc.)
  - NLP (training automatic sem. text processing systems, general information extraction, etc.)
- Extends Tectogrammatical Representation of PCEDT
  - Adds semantic information (for verbs/events)
  - Semantic attributes at each verb occurrence & argument nodes
    - (Synonym) class at verb/predicate/event
      - automatically + manual corrections
    - Semantic roles at arguments
      - automatically (CzEngClass mappings) + manual corrections

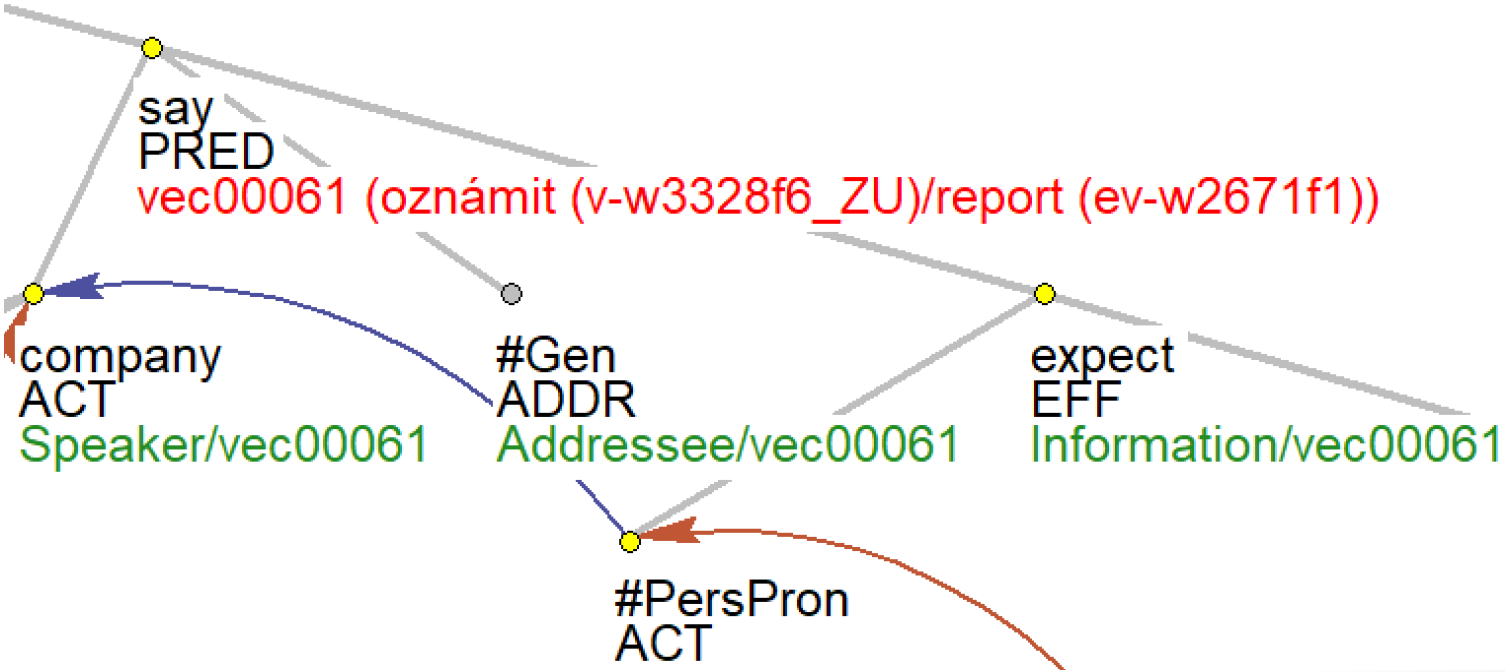
Automatic Pre-annotation

- Coverage of the corpus by the current CzEngClass (100+ classes)
  - 50% independent of alignment
    - En: 67,733 out of 130,079
    - Cz: 48,445 out of 118,029
  - 25% aligned to CzEngClass-covered verb
    - En: 33,005
    - Cz: 32,560
- Up to 5 classes assigned to a single verb occurrence
  - i.e., one valency frame in up to 5 classes
  - 21,050 pairs fully match between the two languages
  - + 5,808 pairs n:n (2:2, 3:3), rest is 1:2, 1:3, 2:1, …

Example: Pre-annotation



Example: Goal (one class only)



Manual Selection and Correction

- Selection of class
  - Substantial number of alignments is *non*-1:1
  - Reasons
    - Verb senses in valency lexicon(s) too coarse-grained: same verb sense in > 1 class
    - Error(s) in class creation
  - Manual selection
    - Disambiguation (of too coarse-grained verb senses)
    - Analysis or errors in CzEngClass
      - Class duplicates → merge the classes
      - Overlapping classes → remove or move some members
        - Create a new class
    - Occasionally error in original PCEDT annotation

Detailed Analysis (Roles) I

- 21 arguments (7,2%) from 100 verb pairs (290 arguments) wrong
- Types of failures on automatic assignment of semantic roles
  - Structural splitting of SR
    - expressing one SR either as one valency argument or two
      - Paul said that he is.PAT-Information a liar. vs.
      - Paul said about him.PAT-Information that he is.EFF-Information a liar.
  - Multiple structural expression of a single SR
    - expressing one SR in multiple syntactic ways not mirrored in the valency frame
      - He.ACT-Speaker called him a liar. vs.
      - In The New York Times.LOC-Speaker, he was called a liar.

Detailed Analysis (Roles) II

- Types of failures on automatic assignment of semantic roles (Cont'd)
  - Reassignment to other nodes, not directly dependent on the verb
    - Role reassigned to more 'deeply' dependent node
      - … expect regulatory.RSTR-Source approval

  - Situational reference
    - newly introduced nodes ("lemma" #SitRef) meant to be linked to the actual situational participant in the current sentence – cannot be automated (so far)
    - similar to textual co-reference (future work)

Conclusions and Future Work

- Conclusions
  - Enrichment of annotated corpus by verb synonym classes
  - Automatic preprocessing insufficient
    - despite same source of information for lexicon creation
    - manual corrections needed
- Future Work
  - Openly available (LINDAT/CLARIAH-CZ repository/service)
  - Comparison with automatic synonym discovery methods
    - Including semi-automatic extensions
  - Use in NLP applications

Thank you!

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

References

- Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The meaning of the sentence in its semantic and pragmatic aspects*. D. Reidel, Dordrecht.
- Zdeňka Urešová, Jan Štěpánek, Jan Hajič, Jarmila Panevová, and Marie Mikulová. 2014. *PDT-Vallex.* LINDAT/CLARIN digital library. http://hdl.handle.net/11858/00-097C-0000-0023-4338-F.
- Zdeňka Urešová, Eva Fučíková, and Jana Šindlerová. 2016. *CzEngVallex: a bilingual Czech-English valency lexicon.* The Prague Bulletin of Mathematical Linguistics, 105:17–50.
- Zdeňka Urešová, Eva Fučíková, Eva Hajičová, and Jan Hajič. 2018a. Creating a Verb Synonym Lexicon Based on a Parallel Corpus. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC'18*), Miyazaki, Japan, May. European Language Resources Association (ELRA).
- Zdeňka Urešová, Eva Fučíková, Eva Hajičová, and Jan Hajič. 2018b. Defining verbal synonyms: between syntax and semantics. In Dag Haug, Stephan Oepen, Lilja Ovrelid, Marie Candito, and Jan Hajič, editors, *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018)* (Pub. No. 155), pages 75–90, Linköping, Sweden. Universitetet i Oslo, Linköping University Electronic Press.
- Zdeňka Urešová, Eva Fučíková, Eva Hajičová, and Jan Hajič. 2018c. Synonymy in Bilingual Context: The CzEngClass Lexicon. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018*, Santa Fe, New Mexico, USA, August 20-26, 2018, pages 2456–2469.