Dvůr Perlová Voda, September 16-18, 2018

Morphology

1 MorFlex -- Jarka
2 Derivational Morphology -- Magda

Valency

3 VALLEX -- Markéta
4 NomVallex -- Veronika
EngVallex -- Silvie
5 PDT-Vallex -- Zdeňka
6 CzEng Vallex -- Zdeňka
7 CzEngClass Lexicon -- Zdeňka

Semantics

Distributional Semantics -- Silvie

Discourse

8 EVALD -- Katka R.
9 AnaConn -- Katka R.
10 Implicit Discourse Relations -- Šárka
11 Lexicon of Discourse Connectives -- Lucka
Complexity in Czech Literary Texts -- Silvie
12 Coreferential chainsin parellel data -- Anja

Treebanks

13 UD -- Dan
14 PDT-C -- Maruška

Treebank Tools

15 PML-TQ -- Matyáš
16 TrEd -- Matyáš
17 KonText -- Pavel S.

Speech

18 Dialog.org -- Nino

Multi-modal Data

19 CEMI -- Pavel P.
20 VIADAT -- Jan
21 ÚSTR -- Jan

Spell-checking

22 Korektor -- Milan

Tagging

23 MorphoDiTa -- Milan

Parsing

24 UDpipe -- Milan
25 Cross-lingual parsing -- Ruda

Named-Entities Recognition

26 NameTag -- Jana

Systems

27 TReex -- Martin
28 UDapi -- Martin

Machine Translation

29 Tensor2Tensor -- Martin
30 Neural Monkey -- Jindra

Sentiment Analysis

31 Katka V.

Chatbots

IBM -- Silvie

Deep Learning

32 LSD -- David

Digital Humanities

Silvie

Infrastructures

33 LINDAT -- Jan
34 DARIAH-CZ -- Jan

INTERNATIONAL PROJECTS

35 Mellon Foundation -- Jan
36 ELG -- Jan
37 SSHOC -- Pavel S.
38 ELITR -- Anja
39 Bergamot -- Anja

ÚFAL

MorFlex

our (=ufalí) morphological dictionary

Recently under reconstruction

The goal

- All the wordforms from PDT (and other our corpora) should be incorporated.
- All the additional information should have exactly the same shape in corpora and in the dictionary.
- Golden rule of morphology

Main changes

- New POS – **FOREIGN WORD**
- Reduction of information about style
- Reduction of explanatory notes within abbreviations
- New conception of variants
  - inflectional variants in the tag
  - global variants interpreted as a sort of derivation
- Adding aspect to the tag (13th position)

Difficult problems

automatic derivations from different sources that lead to the same word

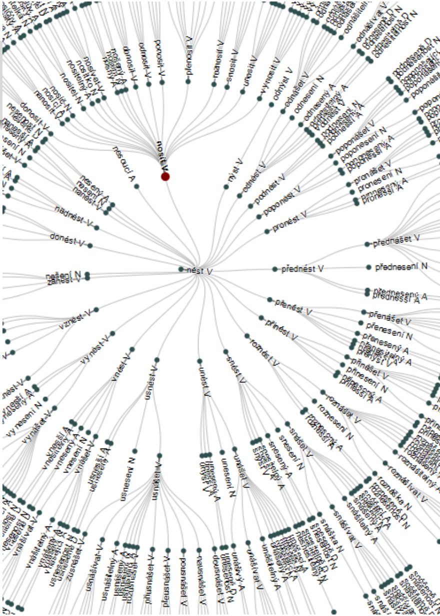
Derivational links?

- two
- no – not morphological problem
- Now – one link, random

Derivational Morphology

People:
Magda Ševčíková, Zdeněk Žabokrtský,
Jonáš Vidra, Lukáš Kyjánek,
Jarmila Panevová, Adéla Kalužová,
Šárka Dohnalová
+ Ruda Rosa

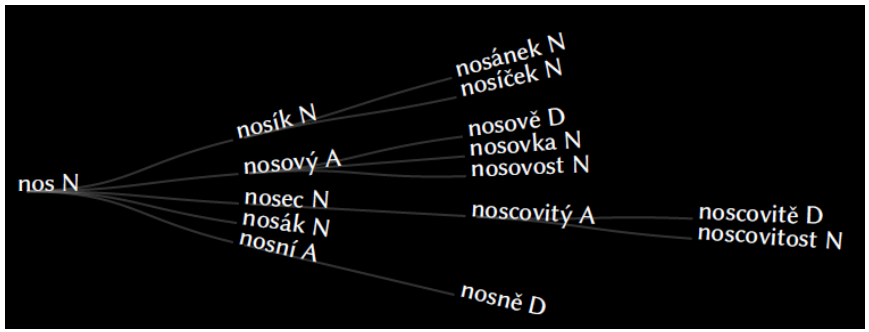
Grants:
GAČR, LINDAT, other grants



DeriNet database

- derivational resource for Czech
  - each word linked to a word which it is supposed to be derived from
  - lemmas from MorfFlex CZ dictionary
- DeriNet 1.5 (2017 in Lindat/Clarin)
  - 1M+ lemmas connected with 785k links
  - 226k lemmas still without a parent
- DeriNet 2.0
  - by the end of 2018
  - new data structure
  - annotation of compounds
- `http://ufal.mff.cuni.cz/derinet`

ID	lemma	POS	parent
384646	nosně (*nasally*)	D	384647
384647	nosní (*nasal*)	A	384650
384648	nosnice (*egg layer*)	N	384654
384649	nosník (*beam, column*)	N	384654
384650	nos (*nose*)	N	
384651	nosnostně ($capacity_{Adv}$)	D	384652
384652	nosnostní ($capacity_{Adj}$)	A	384653
384653	nosnost (*load capacity*)	N	384654
384654	nosný ($bearing_{Adj}$)	A	384625



Recent work

- morphemic analysis using derivational data (Jonáš Vidra)
- pilot experiments on influence of derivational affixes on word embeddings (Zdeněk Žabokrtský et al.)
- collection of existing derivational resources for multiple languages (Lukáš Kyjánek)
- new DeriNet API in Python (Vojtěch Hudeček)
- semi-automatic construction of derivational databases for Polish and Spanish (Mateusz Lango)
- linguistic issues
  - aspect as an inflectional feature expressed by derivation (Jarmila Panevová, MŠ)
  - deverbal nouns vs. denominal verbs in Czech (MŠ)
  - vowel and consonant alternations in derivation (MŠ)
  - modelling derivation of loan words in Czech (MŠ)

Directions to go

- harmonization of resources for multiple languages (Lukáš Kyjánek)
- semi-supervised development of derivational data (incl. deep learning) (Jonáš Vidra)
- annotation and classification of compounds (Adéla Kalužová)
- cross-lingual evolutionary models of derivations (Zdeněk Žabokrtský)
- new linguistic insights
  - paradigms in derivation
  - competition among affixes in word formation
  - correlations between morphemic structure and corpus frequency of words
- Derimo workshop in September 2019 at UFAL

Valency Lexicon of Czech Verbs VALLEX

- provides information on the valency structure of Czech verbs in their particular senses
  - for theoretical linguistic research
  - as an inventory of verb senses (WSD)
  - for Czech learners and users
- phenomena at the lexicon – grammar interface
- theoretical analysis + formal model + application in LR

  goal: to generate all syntactic structure of Czech verbs

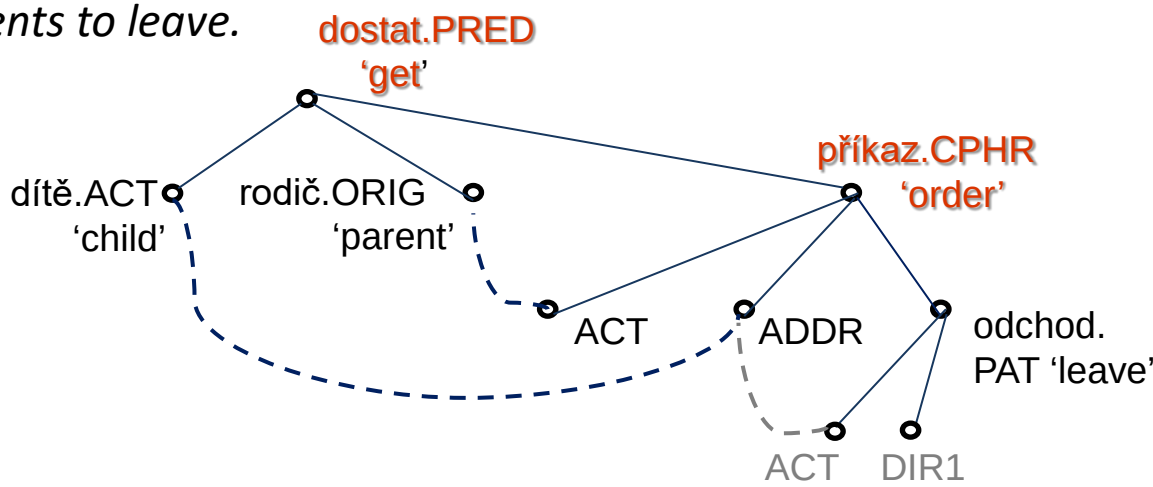
two subprojects:

- light verb constructions as MWEs in VALLEX
- reciprocity and reflexivity

Light Verb Constructions in VALLEX

- two syntactic elements (light verb and predicative noun) serve as a single predicate:
  - predicative noun provides its semantics
  - light verb provides its syntactic capacity
  - coreference as a 'glue'
- syntactically compositional:
  their syntactic structure can be derived by formal rule from the information on light verbs and predicative nouns

*Děti **dostali** od rodičů **příkaz** k odchodu.*
*Children **got** an **order** from (their) parents to leave.*



Reciprocity and Reflexivity in VALLEX

Zabil se *pádem ze střechy.* (CNC, SYN2006pub)
He killed himself (unintentionally) by falling from the roof.

clitic *se* as a part of the verb lemma
→ derived reflexive verbs

Zabil se *vlastní zbraní ...* (CNC, SYN2006pub)/
Zabil *sebe vlastní zbraní.*
He killed himself with his own weapon ...

reflexives *se/sebe* as an object of the verb
→ true reflexives

Zabili *se navzájem. / **Zabili** sebe navzájem.*
They killed each other.

reflexives *se/sebe* as an object of the verb
→ reciprocity

*... **zabila** se dvě vykrmená prasata a pečínka provoněla celý dům.* (CNC, SYN2005)
Two fattened pigs were killed and roast meat scented the whole house.

… and their combinations

reflexives *se* as a part of the verb form
→ deagentive and dispositional diatheses

research tasks:

- analysis from the point of view of theoretical linguistics
- formal model of individual types
- annotation in the data component
- formulation of grammatical rules for grammar component (if applicable)

VALLEX … http://ufal.mff.cuni.cz/vallex/3.0/

award:
2009 – Jednota tlumočníků a překladatelů / Union of Interpreters and Translators:
Slovník roku Contest (3rd place, category *explanatory lexicons*)

to appear: VALLEX 3.5

people:
- Markéta Lopatková
- Vendula Kettnerová
- Anša Vernerová
- Petra Barančíková
- Eda Bejček
- (Zdeněk Žabokrtský)
- Jarmila Panevová (as consultant)

NomVallex

Nominals Valency Lexicon

- **deverbal nouns**
- adjectives
- deadjectival nouns
- …

GAČR project: Corpus-based Valency Lexicon of Czech Nouns

People:

Veronika Kolářová, Jarmila Panevová, Jana Klímová, Anna Vernerová

Duration: 2016-2018 (Oh no!!!)

https://ufal.mff.cuni.cz/grants/nomvallex

NomVallex

Nominals Valency Lexicon

- **deverbal nouns**
- adjectives
- deadjectival nouns
- …

Corpora: CNC SYN, Araneum

Lze však očekávat , že	valence substantiv	, adjektiv a adverbií9 jakožto
ění charakteristické pro	valencí substantiv	, tj. Partitiv) . Když Prouzová
Specifickou otázkou je	valence substantiv	dějových i jmen výsledků děje
řináležitosti patří , totiž	valencí substantiv	, viz Case frames of nouns (
jako centrálního členu .	Valencí substantiv	poměrně podrobně analyzova

GAČR project: Corpus-based Valency Lexicon of Czech Nouns

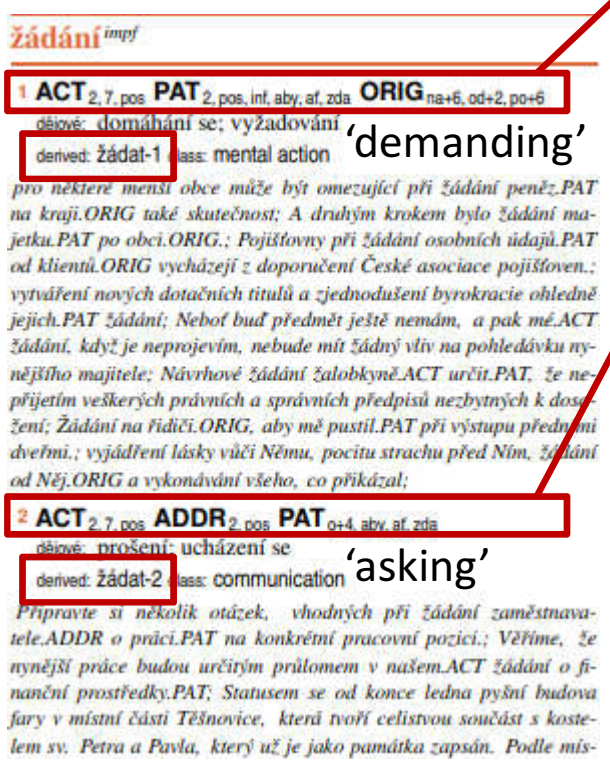
People:

Veronika Kolářová, Jarmila Panevová, Jana Klímová, Anna Vernerová

Duration: 2016-2018 (Oh no!!!)

https://ufal.mff.cuni.cz/grants/nomvallex

NomVallex: Productively derived nouns (-ní/-tí)



žádání *impf*

1 **ACT** 2, 7, pos **PAT** 2, pos, inf, aby, af, zda **ORIG** na+6, od+2, po+6

děj.: domáhání se; vyžadování

derived: žádat-1 class: mental action

pro některé menší obce může být omezující při žádání peněz.PAT na kraji.ORIG také skutečnost; A druhým krokem bylo žádání majetku.PAT po obci.ORIG.; Pojišťovny při žádání osobních údajů.PAT od klientů.ORIG vycházejí z doporučení České asociace pojišťoven.; vytváření nových dotačních titulů a zjednodušení byrokracie ohledně jejich.PAT žádání; Neboť buď předmět ještě nemám, a pak mé.ACT žádání, když je neprojevím, nebude mít žádný vliv na pohledávku nynějšího majitele; Návrhové žádání žalobkyně.ACT určit.PAT, že nepřijetím veškerých právních a správních předpisů nezbytných k dosažení; Žádání na řidiči.ORIG, aby mě pustil.PAT při výstupu předními dveřmi.; vyjádření lásky vůči Němu, pocitu strachu před Ním, žádání od Něj.ORIG a vykonávání všeho, co přikázal;

2 **ACT** 2, 7, pos **ADDR** 2, pos **PAT** o+4, aby, af, zda

děj.: prošení; ucházení se

derived: žádat-2 class: communication

Připravte si několik otázek, vhodných při žádání zaměstnavatele.ADDR o práci.PAT na konkrétní pracovní pozici.; Věříme, že nynější práce budou určitým průlomem v našem.ACT žádání o finanční prostředky.PAT; Statusem se od konce ledna pyšní budova fary v místní části Těšnovice, která tvoří celistvou součást s kostelem sv. Petra a Pavla, který už je jako památka zapsán. Podle mís-

VALLEX: Verbs

žádat *impf*, žádávat *iter*

1 **ACT** 1, inf **PAT** 4, inf, aby, af, zda, že **ORIG** na+6, od+2, po+6

mít požadavky; domáhat se; vyžadovat; chtít

control: ORIG recipr: ACT-ORIG class: mental action ◊ diat: pass deagent

př.: žádat (si.BEN) od někoho omluvu / aby se omluvil; žádal na nás splatit dluh co nejdříve; žádal od něj, zda by mu nepomohl s úkolem; uspět v zaměstnání žádá plné nasazení ◊ pass: *Z dalších programů, které nebývají běžně k dostání, ale podnikatelé jsou dost žádány, stojí za zmínku programy pro výuku jazyků, pro tvorbu cizojazyčné korespondence či překlady do češtiny.* deagent: *žádá se po něm omluva*

2 **ACT** 1 **ADDR** 4 **PAT** o+4, inf, aby, af, zda, že

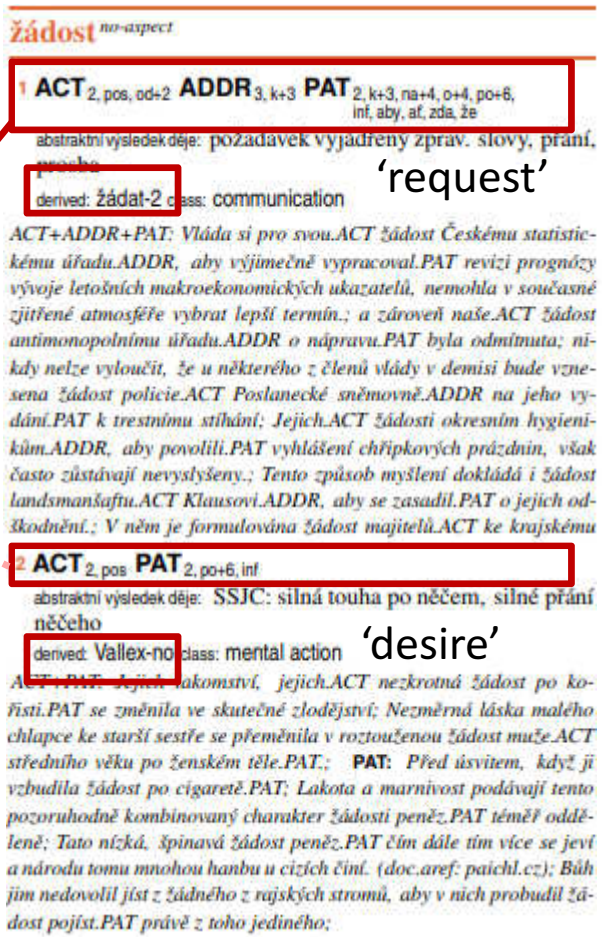
prosit; ucházet se

control: ADDR recipr: ACT-ADDR class: communication ◊ diat: pass deagent

př.: žádat někoho o pomoc / aby se omluvil; Náš mladý soudruh nás v podstatě žádá neměnit výrobu, aby nemusel znovu sčítat poslední kvartál ... (SYN); žádali ho, zda by nechtěl své rozhodnutí odejít ještě přehodnotit ... (SYN) ◊ deagent: *všichni se žádají o pomoc*

3

NomVallex: Non-productively derived nouns (-ba/-ka/-ost/...)



žádost *no-aspect*

1 **ACT** 2, pos, od+2 **ADDR** 3, k+3 **PAT** 2, k+3, na+4, o+4, po+6, inf, aby, af, zda, že

abstraktní výsledek děje: požadavek vyjádřený zprav. slovy, přání, prosba

derived: žádat-2 class: communication

ACT+ADDR+PAT: Vláda si pro svou.ACT žádost Českému statistickému úřadu.ADDR, aby výjimečně vypracoval.PAT revizi prognózy vývoje letošních makroekonomických ukazatelů, nemohla v současné zjitřené atmosféře vybrat lepší termín.; a zároveň naše.ACT žádost antimonopolnímu úřadu.ADDR o nápravu.PAT byla odmítnuta; nikdy nelze vyloučit, že u některého z členů vlády v demisi bude vznesena žádost policie.ACT Poslanecké sněmovně.ADDR na jeho vydání.PAT k trestnímu stíhání; Jejich.ACT žádosti okresním hygienikům.ADDR, aby povolili.PAT vyhlášení chřipkových prázdnin, však často zůstávají nevyslyšeny.; Tento způsob myšlení dokládá i žádost landsmanšaftu.ACT Klausovi.ADDR, aby se zasadil.PAT o jejich odškodnění.; V něm je formulována žádost majitelů.ACT ke krajskému

2 **ACT** 2, pos **PAT** 2, po+6, inf

abstraktní výsledek děje: SSJC: silná touha po něčem, silné přání něčeho

derived: Vallex-no class: mental action

ACT+PAT: Jejich.ACT nezkrotná žádost po kořisti.PAT se změnila ve skutečné zlodějství; Nezměrná láska malého chlapce ke starší sestře se přeměnila v roztouženou žádost muže.ACT středního věku po ženském těle.PAT.; **PAT:** *Před úsvitem, když ji vzbudila žádost po cigaretě.PAT; Lakota a marnivost podávají tento pozoruhodně kombinovaný charakter žádosti peněz.PAT téměř odděleně; Tato nízká, špinavá žádost peněz.PAT čím dále tím více se jeví a národu tomu mnohou hanbu u cizích činí. (doc.aref: paichl.cz); Bůh jim nedovolil jíst z žádného z rajských stromů, aby v nich probudil žádost pojíst.PAT právě z toho jediného;*

'demanding'

'asking'

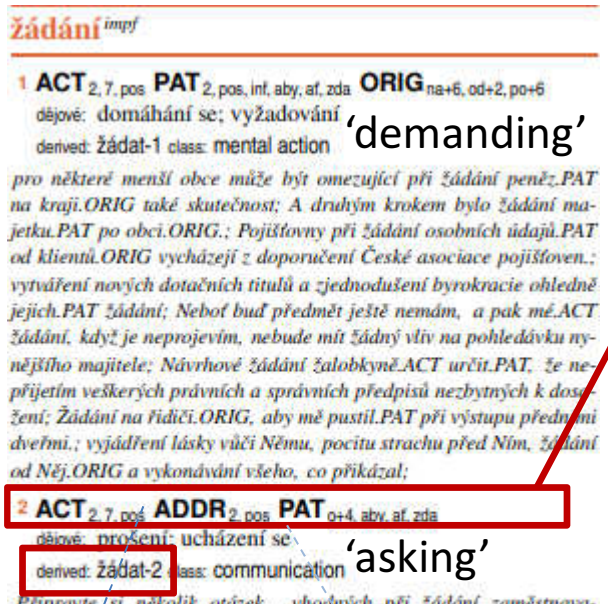
'to demand'

'to ask'

'request'

'desire'

NomVallex: Productively derived nouns (-ní/-tí)



žádání impf

1 ACT 2, 7, pos PAT 2, pos, inf, aby, af, zda ORIG na+6, od+2, po+6
dějově: domáhání se; vyžadování
derived: žádat-1 class: mental action

pro některé menší obce může být omezující při žádání peněz.PAT na kraji.ORIG také skutečnost; A druhým krokem bylo žádání majetku.PAT po obci.ORIG.; Pojišťovny při žádání osobních údajů.PAT od klientů.ORIG vycházejí z doporučení České asociace pojišťoven.; vytváření nových dotačních titulů a zjednodušení byrokracie ohledně jejich.PAT žádání; Neboť buď předmět ještě nemám, a pak mé.ACT žádání, když je neprojevím, nebude mít žádný vliv na pohledávku nynějšího majitele; Návrhové žádání žalobkyně.ACT určit.PAT, že nepřijetím veškerých právních a správních předpisů nezbytných k dosažení; Žádání na řidiči.ORIG, aby mě pustil.PAT při výstupu předními dveřmi.; vyjádření lásky vůči Němu, pocitu strachu před Ním, žádání od Něj.ORIG a vykonávání všeho, co přikázal;

2 ACT 2, 7, pos ADDR 2, pos PAT o+4, aby, af, zda
dějově: prošení; ucházení se
derived: žádat-2 class: communication

Připravte si několik otázek, vhodných při žádání zaměstnavatele.ADDR o práci.PAT na konkrétní pracovní pozici.; Věříme, že nynější práce budou určitým průlomem v našem.ACT žádání o finanční prostředky.PAT; Statusem se od konce ledna pyšní budova fary v místní části Těšnovice, která tvoří celistvou součást s kostelem sv. Petra a Pavla, který už je jako památka zapsán. Podle mís-

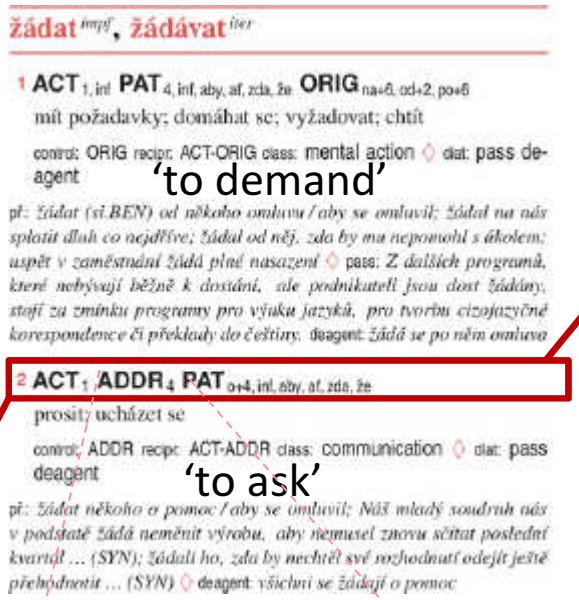
‘demanding’

‘asking’

ADDR: 2,poss

***žádání zaměstnavatele*.ADDR**
***o práci*.PAT**

VALLEX: Verbs



žádat impf, žádávat iter

1 ACT 1, inf PAT 4, inf, aby, af, zda, že ORIG na+6, od+2, po+6
mít požadavky; domáhat se; vyžadovat; chtít

control: ORIG recipr: ACT-ORIG class: mental action ◊ diat: pass deagent

př.: žádat (si.BEN) od někoho omluvu / aby se omluvil; žádat na ně splatit dluh co nejdříve; žádal od něj, zda by mu neponohl s úkolem; uspět v zaměstnání žádá plné nasazení ◊ pass: *Z dalších programů, které nebývají běžně k dostání, ale podnikateli jsou dost žádány, stojí za zmínku programy pro výuku jazyků, pro tvorbu cizojazyčné korespondence či překlady do češtiny.* deagent: *žádá se po něm omluva*

2 ACT 1 ADDR 4 PAT o+4, inf, aby, af, zda, že
prosit; ucházet se

control: ADDR recipr: ACT-ADDR class: communication ◊ diat: pass deagent

př.: žádat někoho o pomoc / aby se omluvil; Náš mladý soudruh nás v podstatě žádá neměnit výrobu, aby nemusel znovu sčítat poslední kvartál ... (SYN); žádali ho, zda by nechtěl své rozhodnutí odejít ještě přehodnotit ... (SYN) ◊ deagent: *všichni se žádají o pomoc*

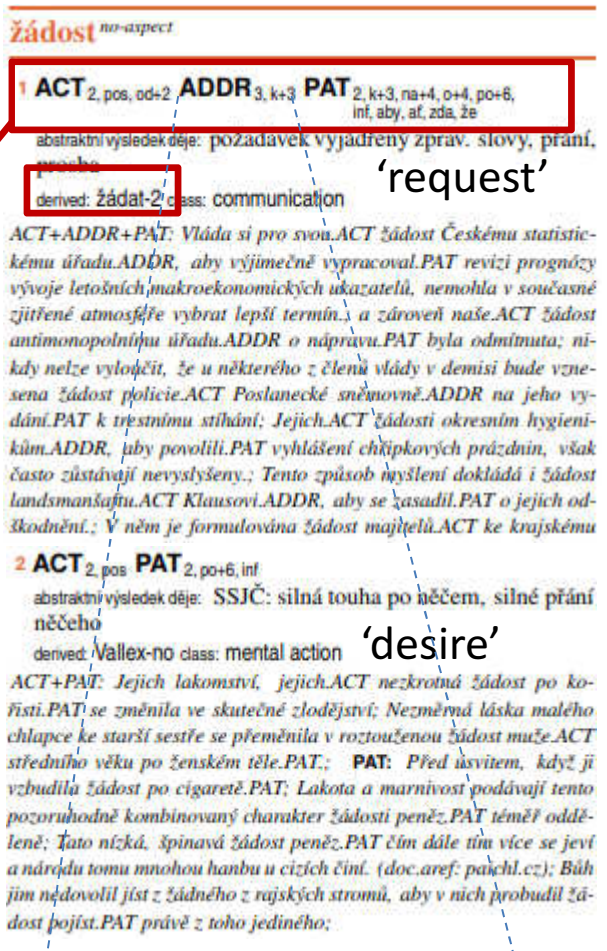
‘to demand’

‘to ask’

ADDR: 4

***žádat zaměstnavatele*.ADDR**
o práci.PAT
‘to ask an employer for a job’

NomVallex: Non-productively derived nouns (-ba/-ka/-ost/...)



žádost no-aspect

1 ACT 2, pos, od+2 ADDR 3, k+3 PAT 2, k+3, na+4, o+4, po+6, inf, aby, af, zda, že
abstraktní výsledek děje: požadavek vyjádřený zprav. slovy, přání, prosba
derived: žádat-2 class: communication

ACT+ADDR+PAT: Vláda si pro svou.ACT žádost Českému statistickému úřadu.ADDR, aby výjimečně vypracoval.PAT revizi prognózy vývoje letošních makroekonomických ukazatelů, nemohla v současné zjitřené atmosféře vybrat lepší termín.; a zároveň naše.ACT žádost antimonopolnímu úřadu.ADDR o nápravu.PAT byla odmítnuta; nikdy nelze vyloučit, že u některého z členů vlády v demisi bude vznesena žádost policie.ACT Poslanecké sněmovně.ADDR na jeho vydání.PAT k trestnímu stíhání; Jejich.ACT žádosti okresním hygienikům.ADDR, aby povolili.PAT vyhlášení chřipkových prázdnin, však často zůstávají nevyslyšeny.; Tento způsob myšlení dokládá i žádost landsmanšaftu.ACT Klausovi.ADDR, aby se zasadil.PAT o jejich odškodnění.; V něm je formulována žádost majitelů.ACT ke krajskému

2 ACT 2, pos PAT 2, po+6, inf
abstraktní výsledek děje: SSJČ: silná touha po něčem, silné přání něčeho
derived: Vallex-no class: mental action

ACT+PAT: Jejich lakomství, jejich.ACT nezkrotná žádost po kořisti.PAT se změnila ve skutečné zlodějství; Nezměrná láska malého chlapce ke starší sestře se přeměnila v roztouženou žádost muže.ACT středního věku po ženském těle.PAT.; **PAT:** *Před úsvitem, když ji vzbudila žádost po cigaretě.PAT; Lakota a marnivost podávají tento pozoruhodně kombinovaný charakter žádosti peněz.PAT téměř odděleně; Tato nízká, špinavá žádost peněz.PAT čím dále tím více se jeví a nárok tomu mnohou hanbu u cizích činí. (doc.aref: pachl.cz); Bůh jim nedovolil jíst z žádného z rajských stromů, aby v nich probudil žádost pojíst.PAT právě z toho jediného;*

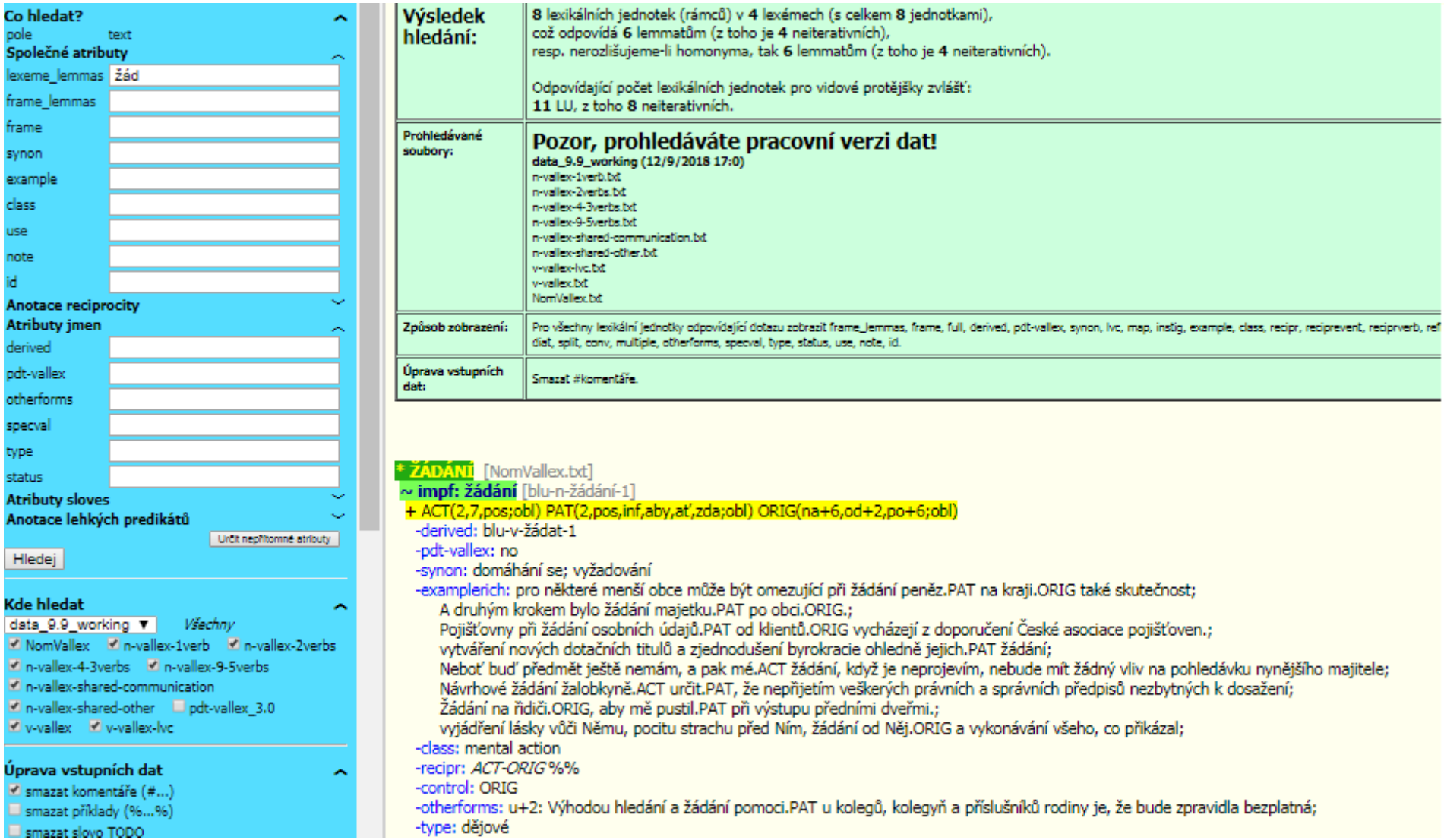
‘request’

‘desire’

ADDR: 3,k+3 ‘to+3’

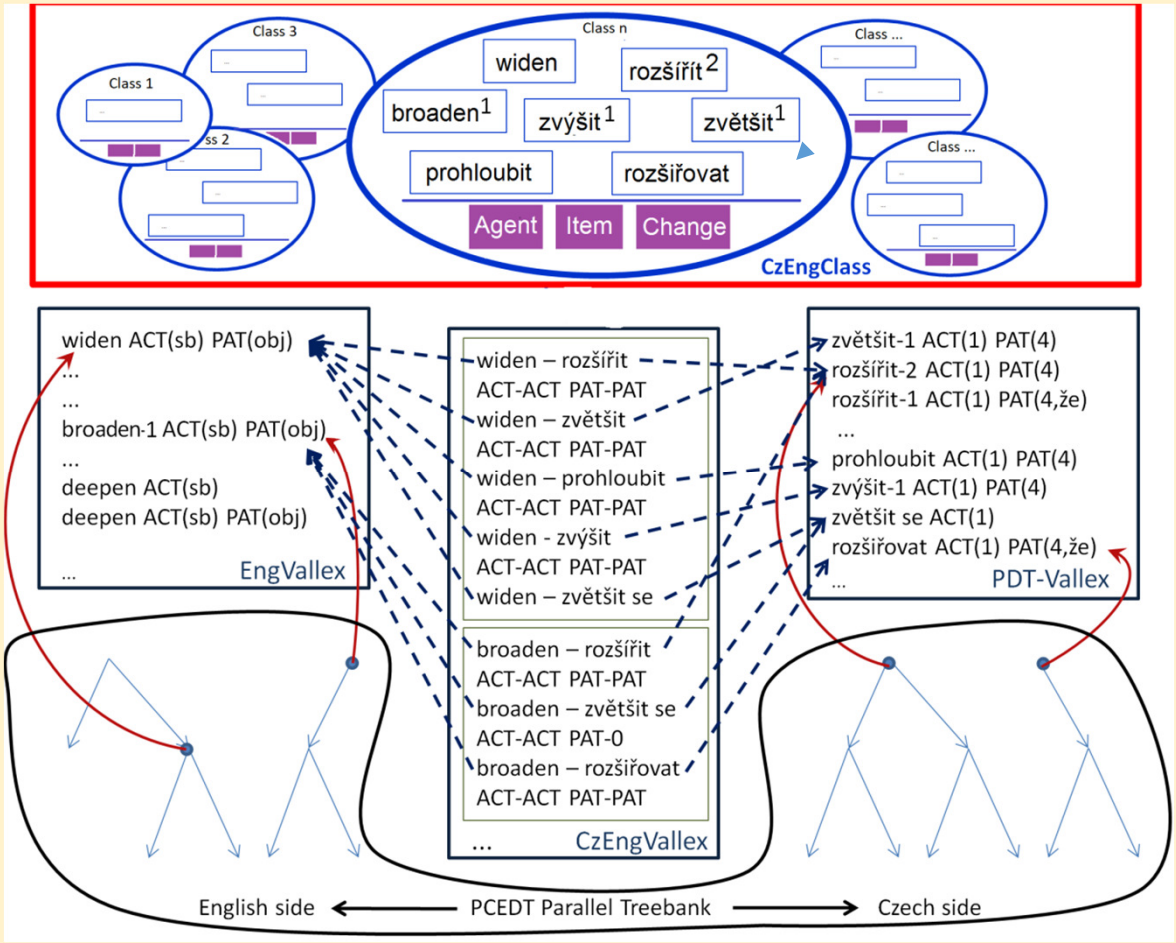
***žádost zaměstnavateli*.ADDR**
***o práci*.PAT**

Search tool: http://quest.ms.mff.cuni.cz/vallex/

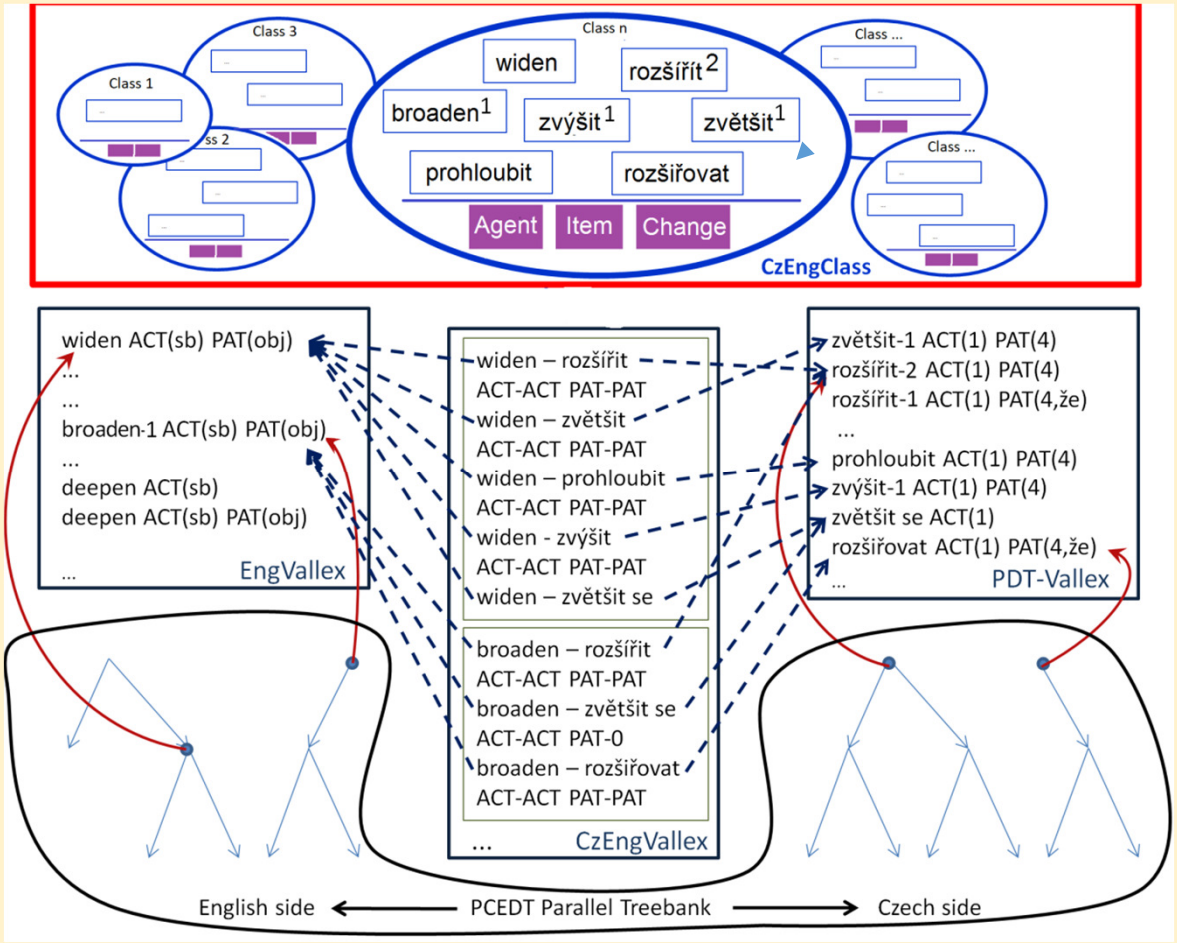


Currently we have about 400 nominal lexical units in NomVallex.

Verbal Synonym Lexicon: CzEngClass

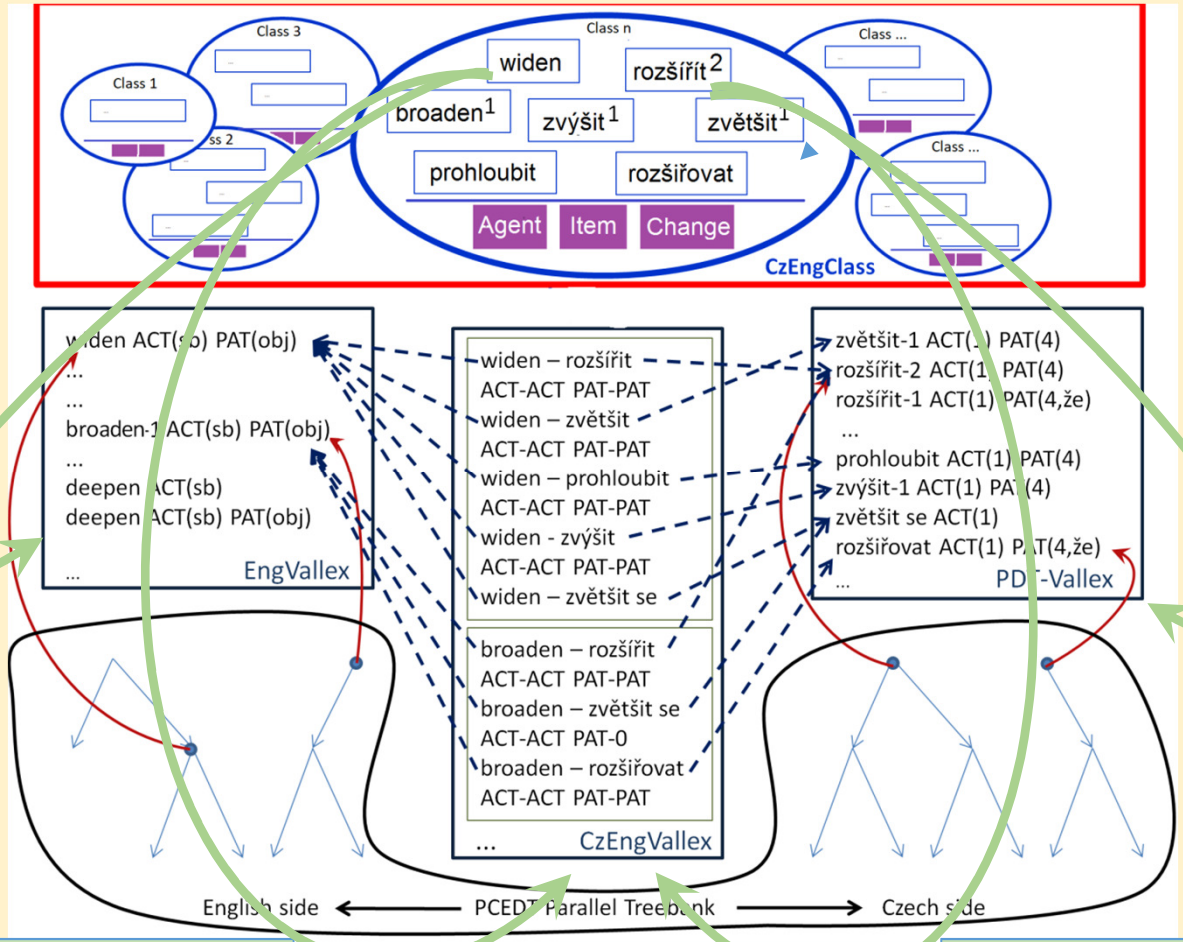


Verbal Synonym Lexicon: CzEngClass



Parallel Corpus

Verbal Synonym Lexicon: CzEngClass



EngVallex

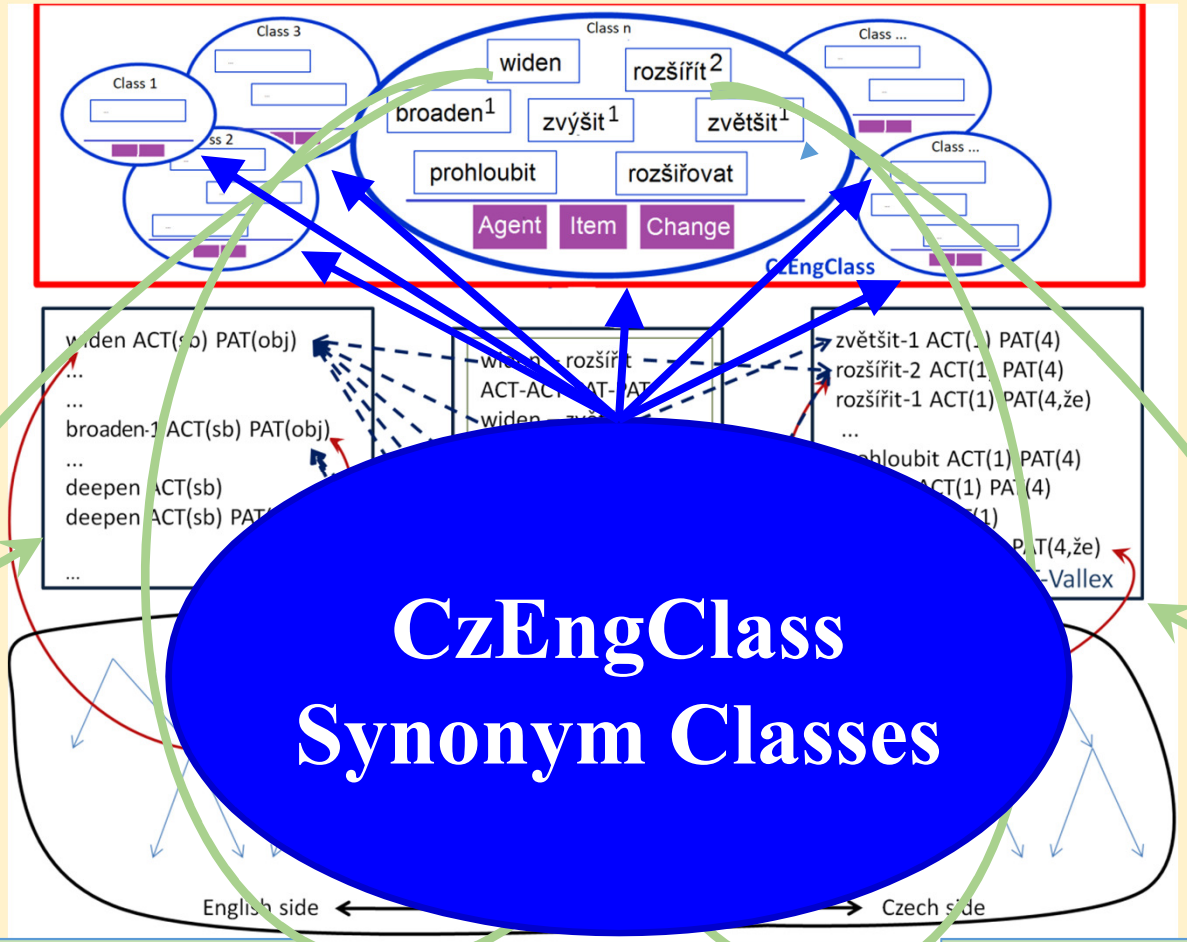
PDT-Vallex

CzEngVallex

Parallel Corpus

CzEngVallex

Verbal Synonym Lexicon: CzEngClass



CzEngClass
Synonym Classes

EngVallex

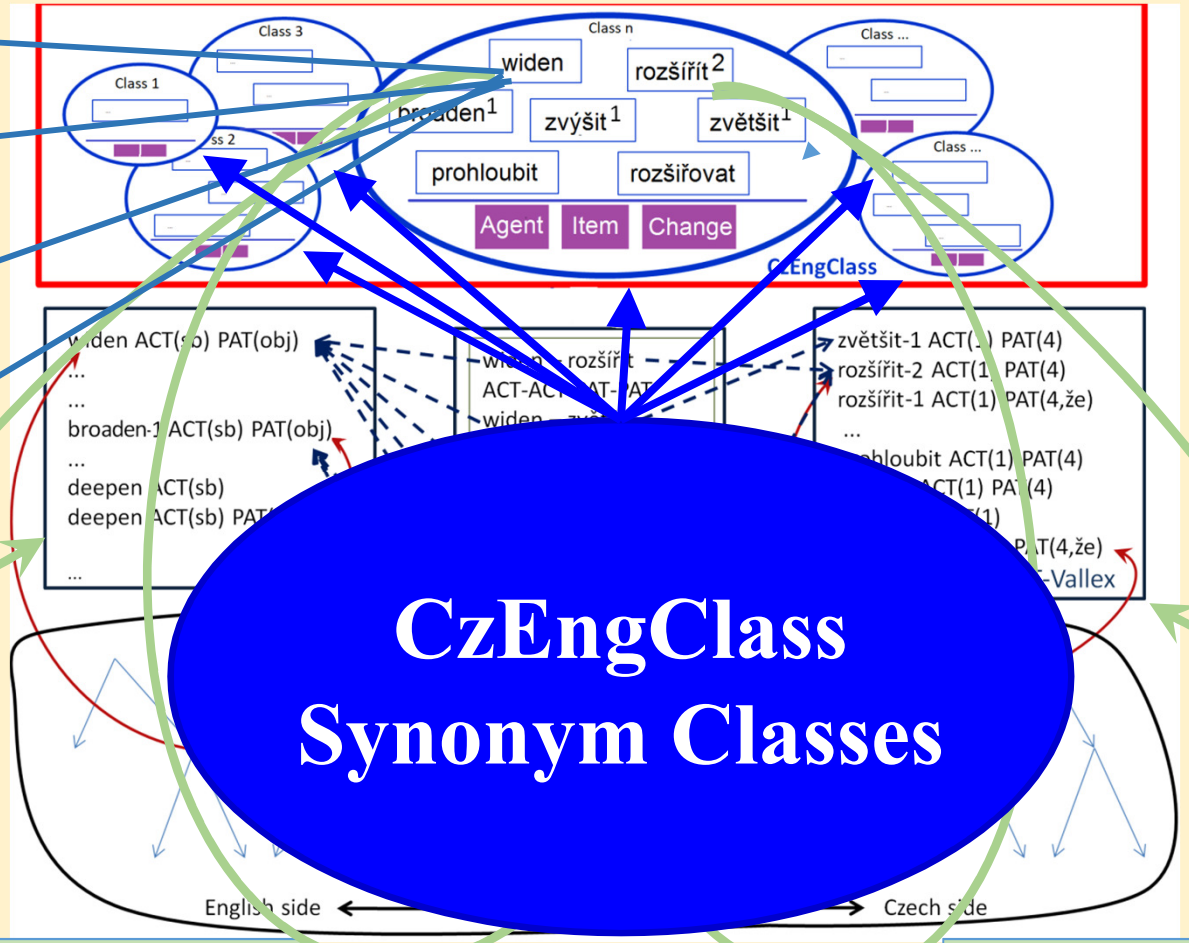
PDT-Vallex

CzEngVallex

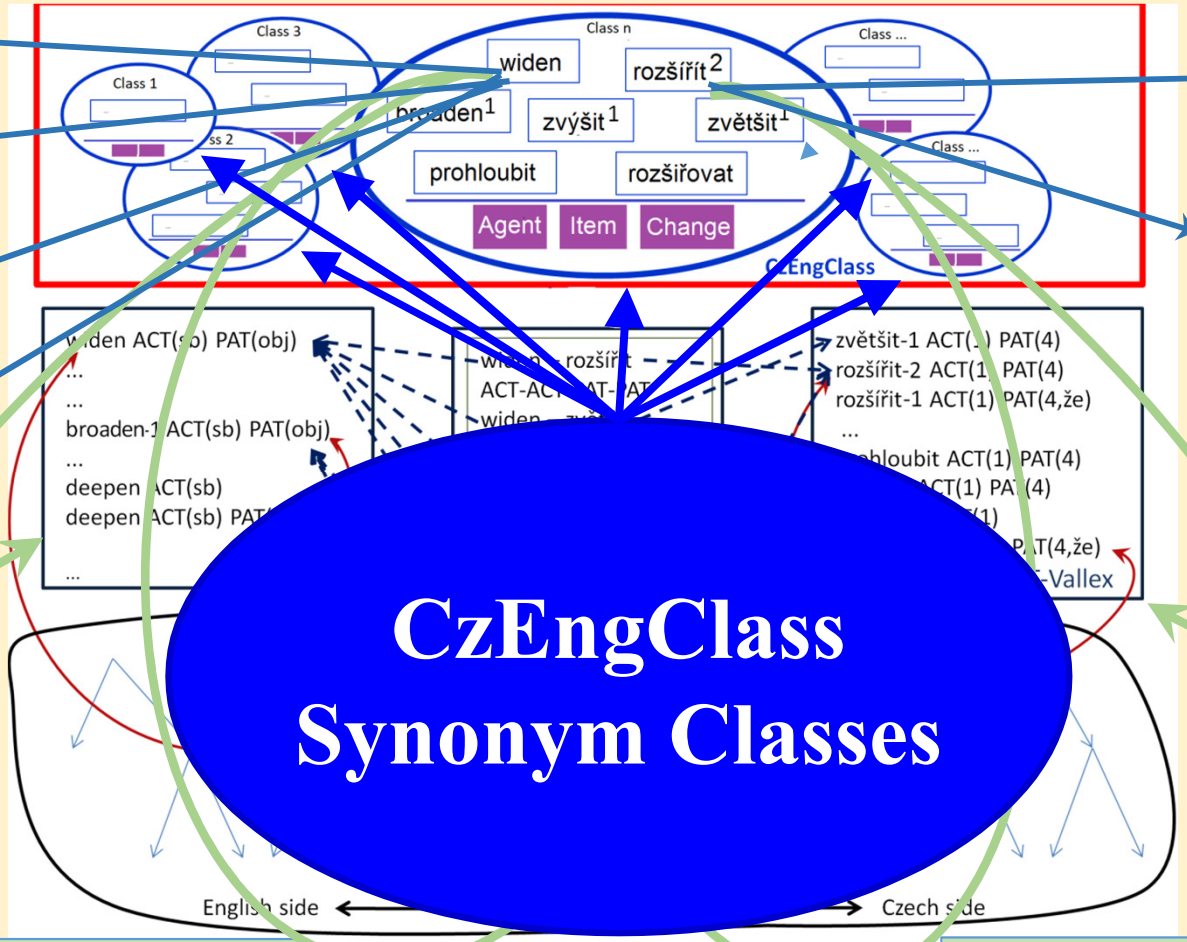
Parallel Corpus

CzEngVallex

Verbal Synonym Lexicon: CzEngClass



Verbal Synonym Lexicon: CzEngClass



FrameNet

WordNet

VerbNet

PropBank

Vallex

Czech WordNet

EngVallex

CzEngClass Synonym Classes

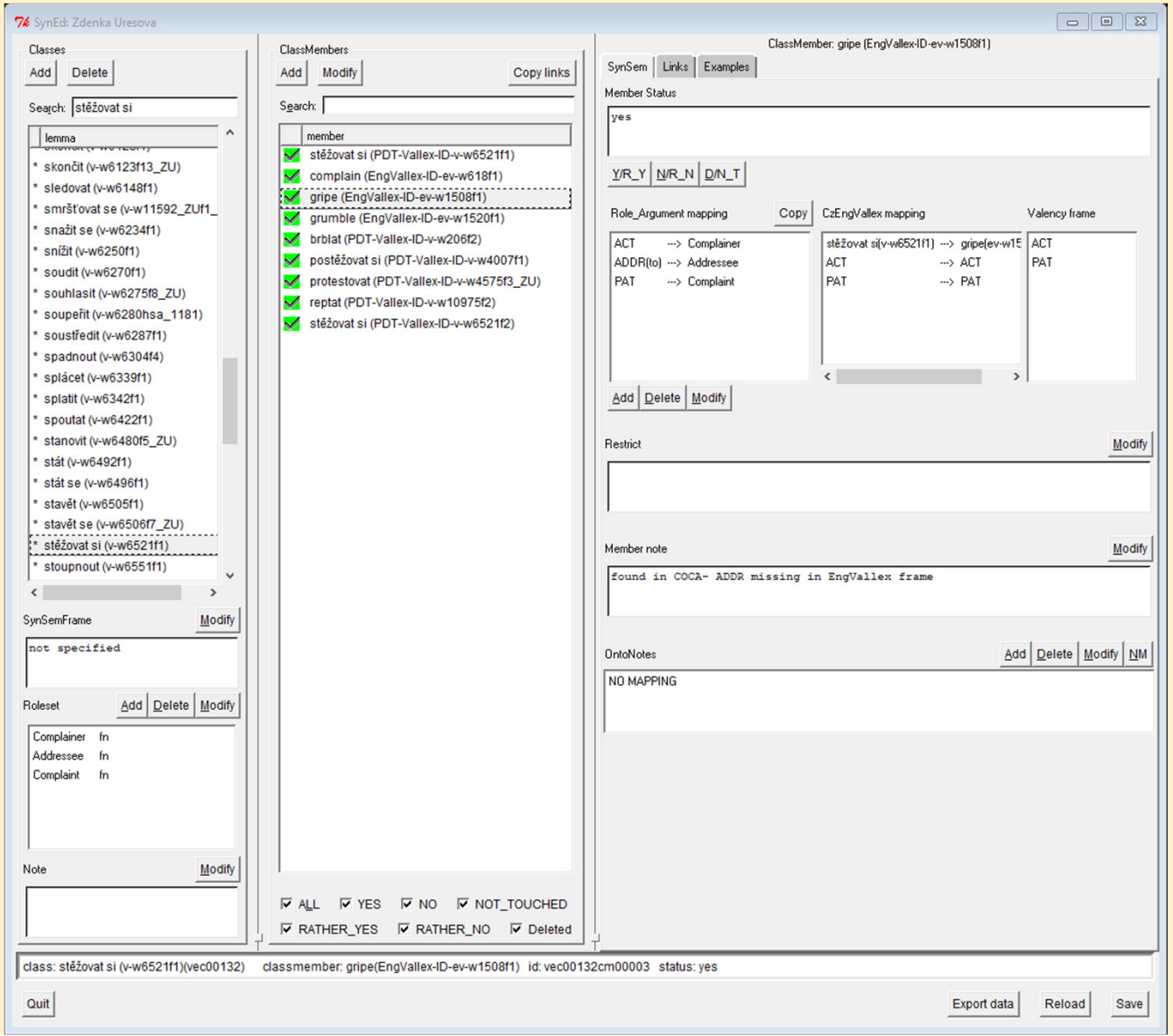
PDT-Vallex

CzEngVallex

Parallel Corpus

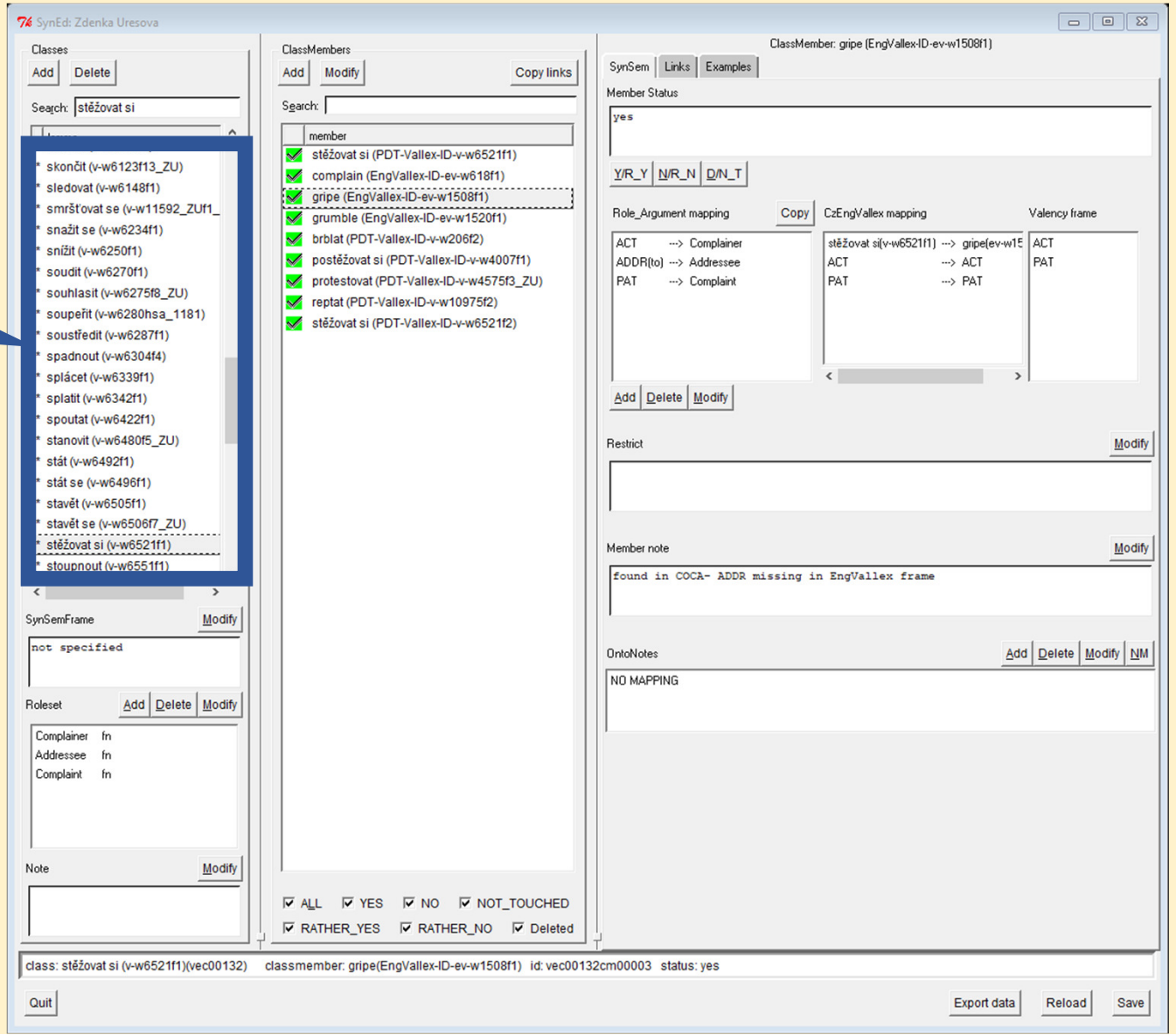
CzEngVallex

SynEd:
Editor for CzEngClass



SynEd:
Editor for CzEngClass

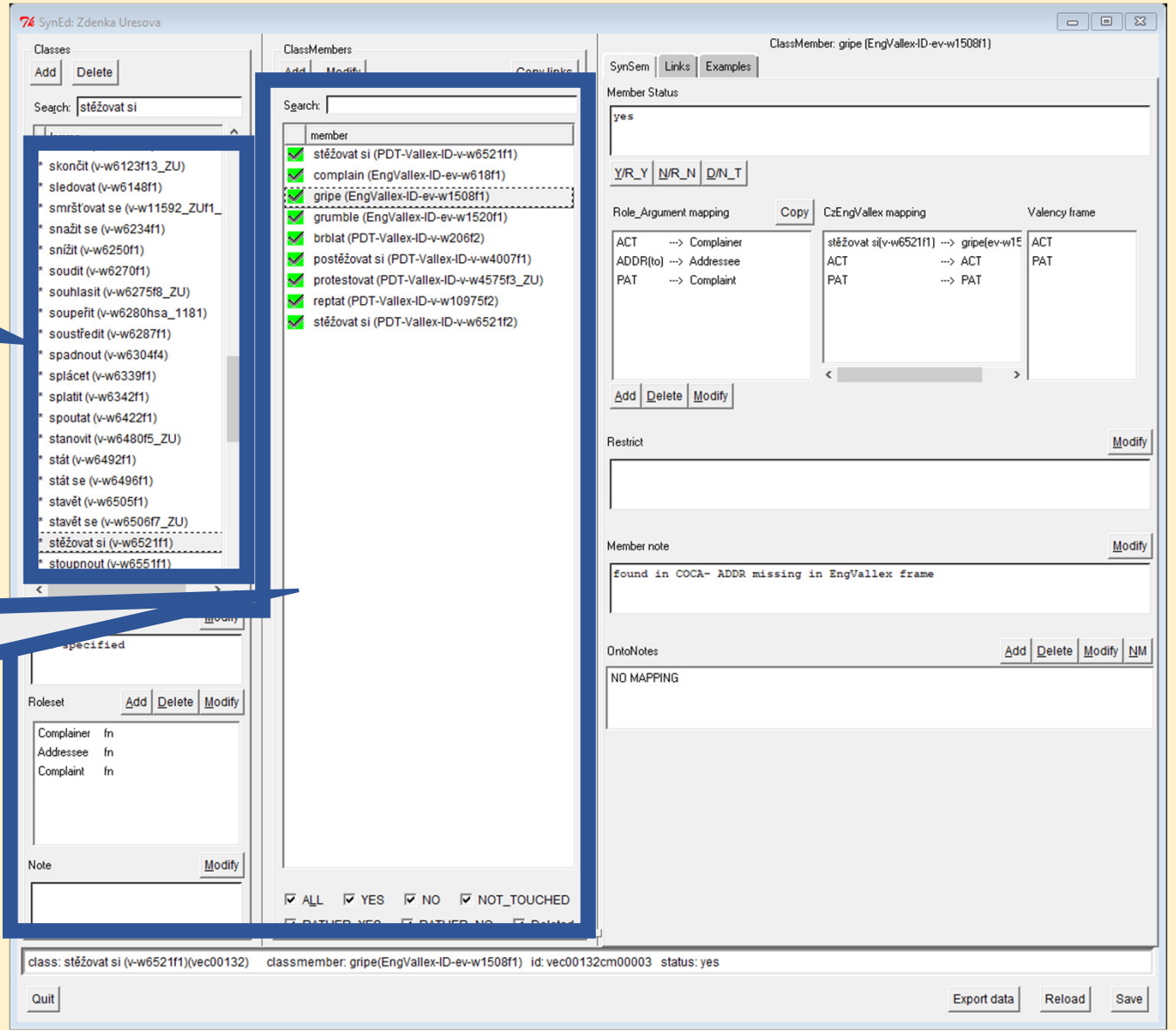
Synonym Classes



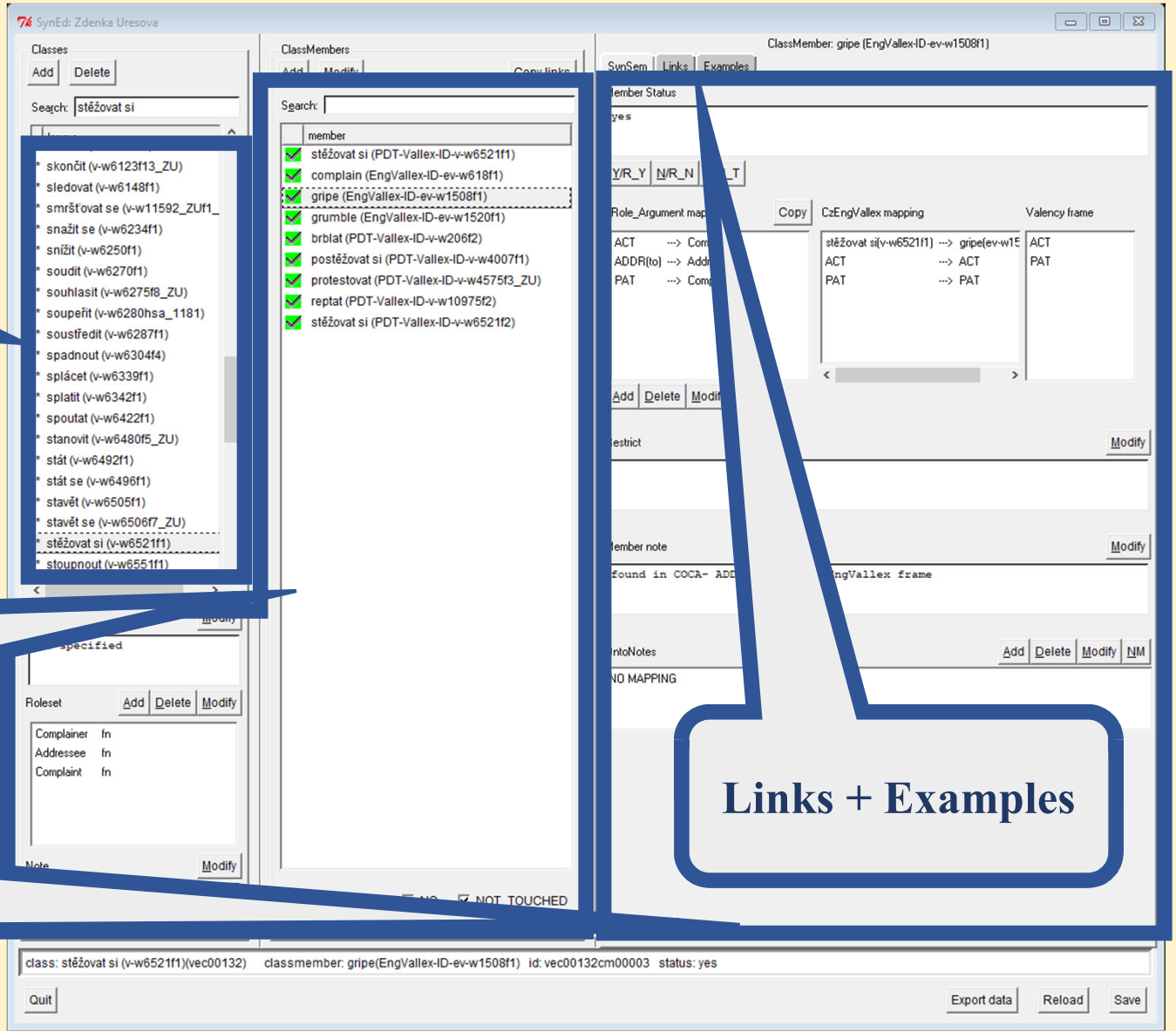
SynEd:
Editor for CzEngClass

Synonym Classes

Editable Info for the Selected Class Labeled *"complain"*



SynEd:
Editor for CzEngClass



Synonym Classes

Editable Info for the Selected Class Labeled *"complain"*

Info about the Class Member selected for editing *gripe*

Links + Examples

Results

- Papers
  - 2017 - Slovko, LTC, Depling
  - 2018 – LREC, Slovanská valence, Coling, Práce filologiczne…TLT?
- LEXICON
  - Openly Aavailable through LINDAT/CLARIN

Urešová, Zdeňka; Fučíková, Eva; Hajičová, Eva; et al., 2018, *CzEngClass 0.2*, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-2824

NAKI II: **Automatic Evaluation of Text Coherence in Czech**

People

Kateřina Rysová, prof. Eva Hajičová, Jiří Mírovský, Michal Novák, Magdaléna Rysová

Main result

Application **EVALD: Evaluator of Discourse**

EVALD – Evaluator of Discourse

- classifier of texts written by **non-native speakers** of Czech (6 categories: from beginners to almost native speakers)

- classifier of texts written by **native speakers** of Czech (5 categories: school marks)

EVALD – Evaluator of Discourse

EVALD classifies text taking into account its:

- **Spelling**: unrecognized words
- **Vocabulary**: complexity and diversity
- **Morphology**: complexity and diversity
- **Syntax**: complexity and diversity
- **Text structure**: frequency of discourse connectives
- **Text structure**: diversity of discourse connectives
- **Text structure**: coreference

EVALD – Evaluator of Discourse

- available also online: **https://lindat.mff.cuni.cz/services/evald-foreign/**

Evaluation

Evaluation class: C1
Probability of the evaluation: 0.39

The text is too short (shorter than 300 words), the evaluation may be inaccurate.

Language aspects stronger than C1:

Spelling: unrecognized words
Vocabulary: complexity and diversity

Language aspects corresponding to C1:

Morphology: complexity and diversity
Text structure: coreference

Language aspects weaker than C1:

Syntax: complexity and diversity
Text structure: frequency of discourse connectives
Text structure: diversity of discourse connectives

GAČR: **Anaphoricity in Connectives: Lexical Description and Bilingual Corpus Analysis**

2017–2019

People

Kateřina Rysová, prof. Eva Hajičová, Jiří Mírovský, Lucie Poláková, Magdaléna Rysová

Main description

- Discourse project
- Connectives in Czech and German

Connectives: small items connecting clauses and sentences

*The blizzard grounded all the flights; **therefore**, she would not be able to fly home for the holidays.*

Anaphoric connectives

- In Czech: ***proto, přitom, přesto, zatímco***…
- In German: ***danach, trotzdem, deswegen***…
- In English: ***therefore, thereby…***

Anaphoric Connectives in Czech

Czech grammaticalized anaphoric connectives	CNC SYN6 (native speakers)	PDiT 2.0 (native speakers)	MERLIN (non-native speakers)
protože "because"	3,207,195	640	259
proto "therefore"	3,190,403	487	90
přitom "while"	1,438,906	261	4
poté "afterwards"	1,253,348	73	1
přesto "yet"	1,131,699	158	2
zatímco "while"	906,394	207	3
potom "then"	653,810	96	37
přestože "though"	564,489	124	4
předtím "before"	381,392	66	3
přičemž while"	293,306	92	0
zato "but still"	166,387	46	0
mezitím "meanwhile"	135,605	32	0
nato "thereafter"	83,026	7	0
natož "let alone"	49,081	14	0
nadto "moreover"	9,981	3	0
mimoto "besides"	7,330	5	0

Anaphoric Connectives in German

Lexeme	DWDS	PCC	MERLIN	Lexeme	DWDS	PCC	MERLIN	Lexeme	DWDS	PCC	MERLIN
damit	70,788	45	55	**nachher**	4,806	0	2	**hiermit**	1,620	0	16
dabei	47,321	20	14	worauf	4,396	0	2	**demgegenüber**	1,601	0	0
dazu	43,693	16	38	**seitdem**	4,309	1	2	**hierdurch**	1,283	0	0
darauf	43,266	12	29	**deswegen**	3,721	1	29	**stattdessen**	1,190	0	1
dafür	29,092	19	50	**hingegen**	3,639	2	3	**mithin**	1,189	1	0
daher	25,460	3	11	wonach	3,497	1	0	**unterdessen**	865	0	0
dadurch	23,319	2	6	wodurch	3,427	0	0	wogegen	724	0	0
indem	19,712	4	0	**zudem**	3,209	4	0	**demzufolge**	611	0	0
dagegen	19,468	7	4	womit	3,042	0	1	hiernach	482	0	0
nachdem	18,835	4	11	davor	2,872	0	1	**nebenher**	425	0	0
darum	15,960	4	11	**daraufhin**	2,783	1	1	weswegen	349	0	0
wobei	15,336	1	0	**hierfür**	2,780	0	0	**währenddessen**	337	0	0
danach	11,936	1	19	**demnach**	2,612	0	0	**woraufhin**	199	0	0
vorher	11,659	2	5	**überdies**	2,560	0	0	**wohingegen**	138	0	0
trotzdem	10,153	3	19	**infolgedessen**	2,237	0	0	**dahingegen**	35	0	0
hierzu	7,558	0	0	**hierauf**	2,086	0	0	**dementgegen**	2	0	0
indessen	6,193	1	0	**hinterher**	1,853	0	0	**hieraufhin**	1	0	0
daneben	5,170	0	0	**seither**	1,782	1	1				

Implicit discourse relations

GA ČR project (2017-2019)

Šárka Zikánová, Jiří Mírovský,
Pavlína Synková

Za předchozího režimu si jazzoví fandové lépe uvědomovali výjimečnost těchto akcí [jazzových koncertů].

Dnes si mnozí říkají, že i hvězdy mohou vidět jindy

In the previous regime, jazz fans were more aware of the uniqueness of these events [jazz concerts].

Today, many say that even the stars can be seen at other times.

Za předchozího režimu si jazzoví fandové lépe uvědomovali výjimečnost těchto akcí [jazzových koncertů].

Confrontation

Dnes si mnozí říkají, že i hvězdy mohou vidět jindy

In the previous regime, jazz fans were more aware of the uniqueness of these events [jazz concerts],

Confrontation

Today, many say that even the stars can be seen at other times.

Linguistic research on the text structure

- Means establishing text coherence in Czech
  - Formal means: verbal aspect and tense, temporal settings, modal verbs, typical reflection of the information structure in the surface word order, specific syntactic structures
  - World knowledge: relations between entities and processes (hyperonymy, co-hyponymy, implication etc.)

Typical patterns:

hyperonymy – specification / generalization

co-hyponymy – conjunction / confrontation

evaluation in the focus – explication in the next sentence

- Distribution of explicit and implicit discourse relations at different semantic types of relations

- Interplay between the identified means establishing text coherence and other features of text (complexity of the syntactic structure, text genre, expressed / omitted subject etc.)

- Annotation: 5000 sentences
- Experiments on the explicitness / implicitness of discourse relations

CzeDLex 0.5

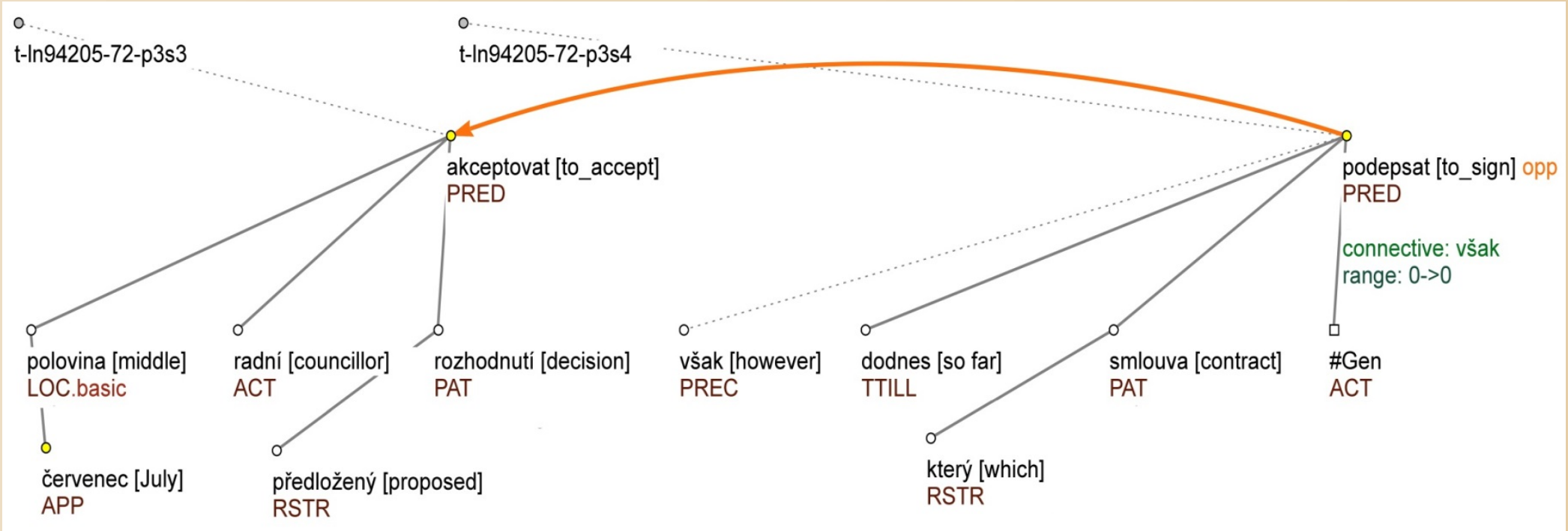
Lexicon of Czech Discourse Connectives

J. Mírovský, P. Synková, M. Rysová, L. Poláková

Discourse relations and connectives in PDT

V polovině července radní předložené rozhodnutí akceptovali. **Dodnes však žádná smlouva podepsána nebyla.** **(vztah opozice)**

[*In middle July, the councilors accepted the proposed decision.* **However, no contract was signed so far.**] **(disc. relation of opposition)**



CzeDLex 0.5

published on December 24, 2017

MŠMT Cost-cz project (2015–2017)

PDT 3.5: **21 223** manually annotated discourse relations:

-> **automatic** extraction of connectives into a lexicon
& **manual** checks and additions

-> **205** entries (different connectives), **18** post-processed manually
(> 2/3 of all discourse relations in the data)

-> **secondary connectives** included (e.g. *z toho důvodu/ for this reason; v případě, že/ in case that…*)

J. Mírovský, P. Synková, M. Rysová, L. Poláková: ***CzeDLex 0.5.*** Data/software, Univerzita Karlova, Česká republika, http://hdl.handle.net/11234/1-2538, 2017

CzeDLex 0.5

published on December 24, 2017

MŠMT Cost-cz project (2015–2017)

- CzeDLex in Tred (editing) and on the web (searching, filters)
- This year no founding, but:
  - work on incorporation to multilingual Connective-Lex (http://connective-lex.info; 7 languages so far)
  - GAČR proposal for 2019-2021 to complete CzeDLex (mainly manually) and use it as a compoment of a future discourse parser
- Byproduct – manual English translations of PDT ☺

J. Mírovský, P. Synková, M. Rysová, L. Poláková: ***CzeDLex 0.5.*** Data/software, Univerzita Karlova, Česká republika, http://hdl.handle.net/11234/1-2538, 2017

CzeDLex 0.5

basic | discourse types | parts of speech | documentation

all
concession
condition
confrontation
conjunction
conjunctive alternative
correction
disjunctive alternative
equivalence
explication
generalization
gradation
instantiation
opposition
pragmatic condition
pragmatic contrast
pragmatic reason-result
precedence-succession
purpose
reason-result
restrictive

odůvodnění
ohled
okamžik
oproti
ostatně
ovšem [but, of course]
pak [then]
plynout
podmínka
podobně
pokračovat
pokud [if]
poněvadž
popřípadě
poslední
posléze
potom
potom co
potom kdy
poté
poté co
poté kdy
pouze
později
pravděpodobnější
prostě
proto [therefore]
protože [because]
první

protože [because] (primary, single; count: 640)

variants: proto že / proto že proto že / proto že že / že proto že

connective usages (99%; intra 98%)

reason-result (poněvadž [because], 98%; intra 99%; subordinating conjunction)

Note: inter-sentential use = parcelling of the dependent clause
[arg_semantics: reason-result:reason; ordering: any; integration: first]
complex_forms (0%): protože protože [because because] (mult; two coordinated because-clauses) / protože tak [because so] (corr)
modifications (0%): právě proto že [exactly because]
examples:

Lidé byli spokojeni, protože si více vydělali.
[People were happy because they made more money.]
Neměli na to čas, protože byli většinou v terénu.
[They did not have time for it because they were mostly in the field.]
J. Odložil upadl, protože jej někdo fyzicky napadl.
[J. Odložil fell down because someone had physically assaulted him.]
A nevysílají české Události právě pro ty banality. Protože právě jejich znalost by mohla na Slovensku dělat neplechu.
[And they do not broadcast Czech Události namely for those banalities. Because namely their knowledge could cause troubles in Slovakia.]

explication (totiž [because], 1%; intra 67%; subordinating conjunction)

[arg_semantics: explication:argument; ordering: any; integration: first]
examples:

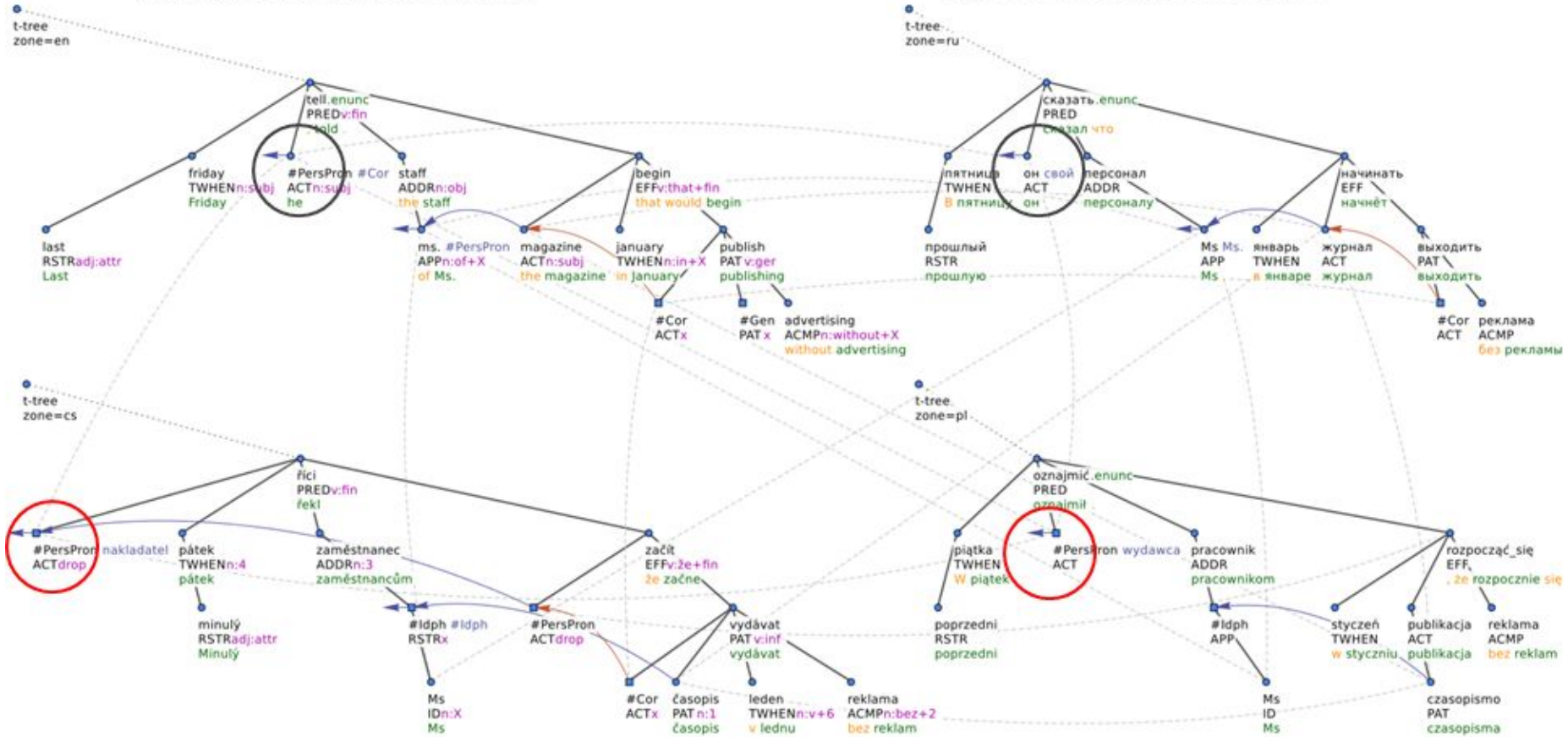
Anebo, jak napsal jistý izraelský komentátor, rozhodl se urazit se.

Structure of coreferential chains in parallel language data (Anja Nedoluzhko, Michal Novák)

- *What it is about*: coreference relations in parallel data, comparison of coreference between languages, towards coreference-based language typology, coreference resolution (CR) for different languages, annotated data for linguistic analysis and coreference resolution
- *Results*:
  - Linguistic: Contrastive analyses (reflexive possessives, zeros, pronominal adverbs, correlative constructions, personal vs. impersonal constructions)
  - Computational: using cross-lingual methods (bilingually informed CR, coreference projection) to quantify differences
  - Data: PCEDT_Coref → PAWS

EN: Last Friday he told the staff of Ms. that the magazine in January would begin publishing without advertising.

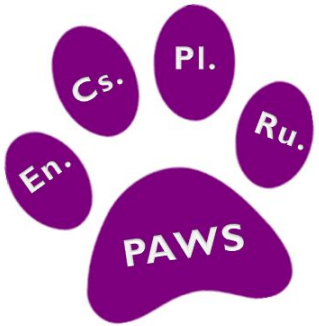
RU: В прошлую пятницу он сказал персоналу Ms, что в январе журнал начнёт выходить без рекламы.



CS: Minulý pátek řekl zaměstnancům Ms., že časopis v lednu začne vydávat bez reklam.

PL: W poprzedni piątek oznajmił pracownikom Ms., że w styczniu publikacja czasopisma rozpocznie się bez reklam.

PAWS: Parallel Anaphoric Wall Street Journal



- A first half of the PCEDT section 19, particularly the 50 documents from wsj 1900 to wsj 1949
- Translated into Russian and Polish
- Manual annotation of word alignment

	English	Czech	Russian	Polish
Sentences		1,078		
Tokens	26,149	25,697	25,704	25,763
Tectogrammatical nodes	18,611	20,696	18,874	18,541
Coreferring nodes	4,210	4,403	4,254	3,371
grammatical coreference	729	528	749	294
textual pron. coref. overt	544	213	493	206
textual pron. coref. elided	76	643	32	243
textual nominal coreference	1,361	1,496	1,610	1,568
first mentions	1,277	1,330	1,243	979
reference to split antecedents	149	149	91	65
reference to a segment	28	23	16	12
exophora	46	21	20	4

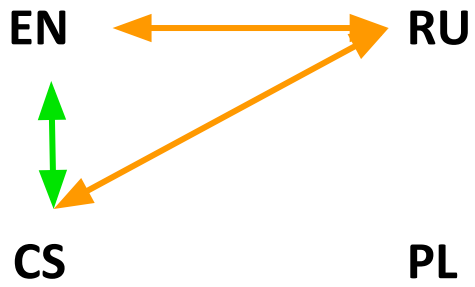
Linguistic typology using computational methods

Characteristics based on two different methods for parallel texts:

BICR: *F(*bilingually informed CR*)* - *F(*monolingual CR*)*

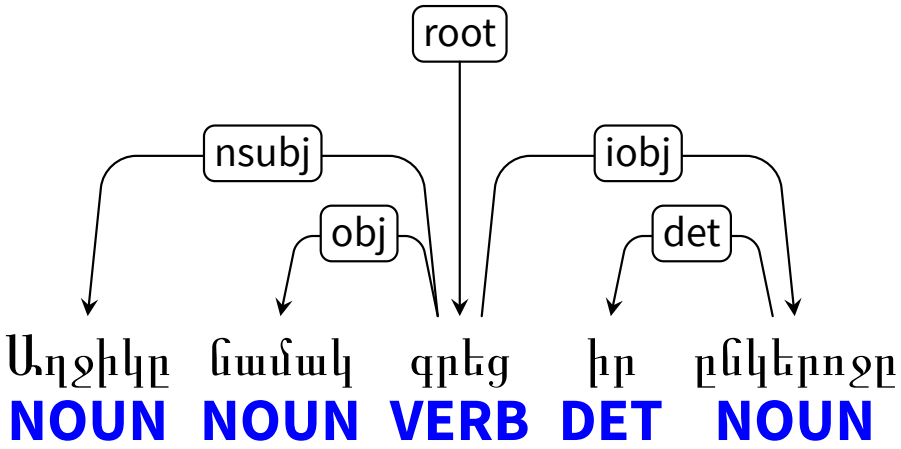
PROJ: *F(*CR trained on gold projections*)* - *F(*gold projections*)*

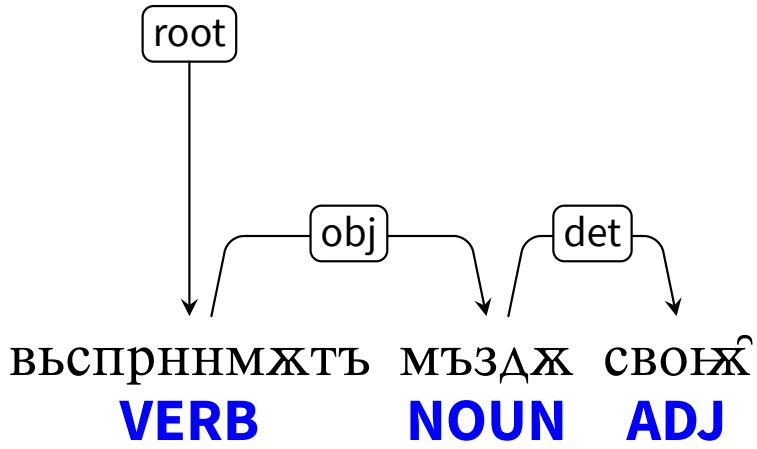
	CS (with EN)	EN (with CS)
BICR	+1.9	+1.5
PROJ	-12.4	-20.8

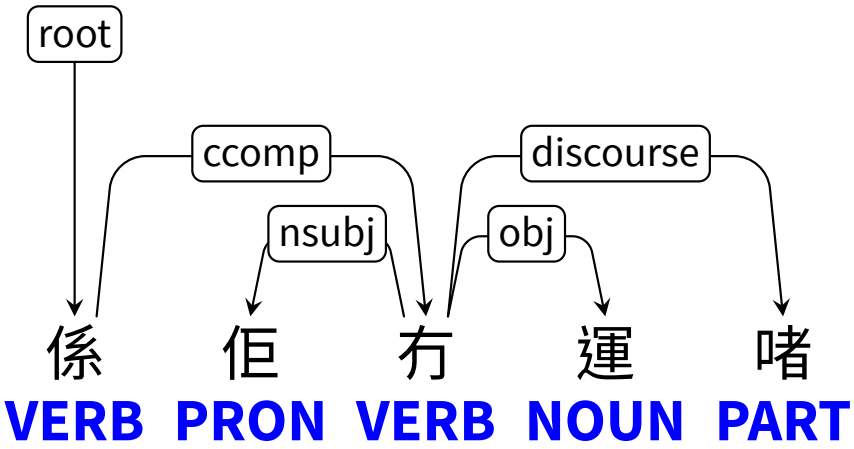


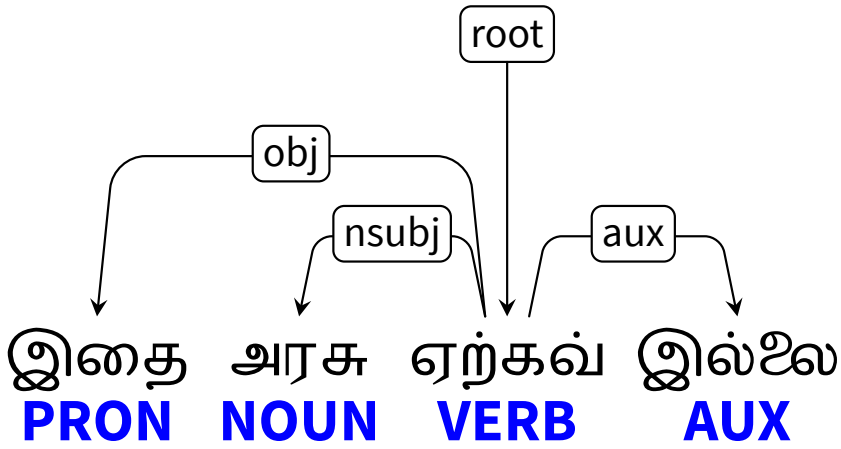
English seems to be more informative for Czech than vice versa.

Universal Dependencies









A Big Family: 71 Languages and Growing

- I.-E.: Armenian, Ancient Greek, Greek, Breton, Irish
  - Germanic: Afrikaans, Danish, Dutch, English, Faroese, German, Gothic, Norwegian, Swedish
  - Romance: Catalan, French, Galician, Italian, Latin, Old French, Portuguese, Romanian, Spanish
  - Balto-Slavic: Belarusian, Bulgarian, Croatian, Czech, Church Slavonic, Polish, Russian, Serbian, Slovak, Slovenian, Ukrainian, Sorbian, Latvian, Lithuanian
  - Indo-Ir.: Kurmanji, Persian, Hindi, Marathi, Sanskrit, Urdu
- Uralic: Estonian, Finnish, Hungarian, Komi, Sami
- Turkic: Kazakh, Turkish, Uyghur
- Dravidian: Tamil, Telugu
- Afro-Asiatic: Amharic, Arabic, Coptic, Hebrew
- Sino-Tibetan: Cantonese, Chinese
- Austro-Asiatic: Vietnamese; Tai-Kadai: Thai
- Austronesian: Indonesian, Tagalog
- Other: Buryat, Japanese, Korean, Basque, Sw. Sign, Naija, Yoruba, Warlpiri

266 Contributors

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, John Bauer, Sandra Bellato, Kepa Bengoetxea, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Rogier Blokland, Victoria Bobicev, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Jayeol Chun, **Silvie Cinková,** Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Carly Dickerson, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, **Kira Droganova, Puneet Dwivedi,** Marhaba Eli, Ali Elkahky, Binyam Ephrem, Tomaž Erjavec, Aline Etienne, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Céline. Guillot-Barbance, Nizar Habash, **Jan Hajič, Jan Hajič jr.,** Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, **Barbora Hladká, Jaroslava Hlaváčová,** Florinel Hociung, Petter Hohle, Jena Hwang, Radu Ion, Elena Irimia, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Tolga Kayadelen, **Václava Kettnerová,** Jesse Kirchner, Natalia Kotsyba, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Tatiana Lando, **Septina Dian Larasati,** Alexei Lavrentiev, John Lee, Phương Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, KyungTae Lim, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, **David Mareček,** Katrin Marheinecke, Héctor Martínez Alonso, André Martins, **Jan Mašek,** Yuji Matsumoto, Ryan McDonald, Gustavo Mendonça, Niko Miekka, Anna Missilä, Cătălin Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Shinsuke Mori, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, **Anna Nedoluzhko,** Gunta Nešpore-Bērzkalne, Lương Nguyễn Thị, Huyền Nguyễn Thị Minh, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, **Adédayọ̀ Olúòkun,** Mai Omura, Petya Osenova, Robert Östling, Lilja Øvrelid, Niko Partanen, **Elena Pascual,** Marco Passarotti, Agnieszka Patejuk, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Emily Pitler, Barbara Plank, Thierry Poibeau, **Martin Popel,** Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Andriela Rääbis, Alexandre Rademaker, **Loganathan Ramasamy,** Taraka Rama, Carlos Ramisch, **Vinit Ravishankar,** Livy Real, Siva Reddy, Georg Rehm, Michael Rießler, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, **Rudolf Rosa,** Davide Rovati, Valentin Roșca, Olga Rudina, Shoval Sadde, Shadi Saleh, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Baiba Saulīte, Yanin Sawanakunanon, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Muh Shohibussirri, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Isabela Soares-Bastos, Antonio Stella, **Milan Straka,** Jana Strnadová, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Yuta Takahashi, Takaaki Tanaka, Isabelle Tellier, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, **Zdeňka Urešová,** Larraitz Uria, Hans Uszkoreit, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Veronika Vincze, Lars Wallin, Jonathan North Washington, Seyi Williams, Mats Wirén, Tsegay Woldemariam, Tak-sum Wong, Chunxiao Yan, Marat M. Yavrumyan, Zhuoran Yu, **Zdeněk Žabokrtský,** Amir Zeldes, **Daniel Zeman,** Manying Zhang, Hanzhi Zhu

Current ÚFAL Involvement

- Core UD group, guidelines, releases  

Current ÚFAL Involvement

- Core UD group, guidelines, releases  
- CoNLL 2017 & 2018 parsing shared tasks    

Current ÚFAL Involvement

- Core UD group, guidelines, releases  
- CoNLL 2017 & 2018 parsing shared tasks    
- Annotation, conversion    

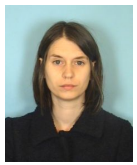
Current ÚFAL Involvement

- Core UD group, guidelines, releases  
- CoNLL 2017 & 2018 parsing shared tasks    
- Annotation, conversion    
- Tools   

Current ÚFAL Involvement

- Core UD group, guidelines, releases  
- CoNLL 2017 & 2018 parsing shared tasks    
- Annotation, conversion    
- Tools   
- Online query  

Current ÚFAL Involvement

- Core UD group, guidelines, releases  
- CoNLL 2017 & 2018 parsing shared tasks    
- Annotation, conversion    
- Tools   
- Online query  
- Linguistic research   Parsing  

Prague Dependency Treebank Consolidated PDT-C 1.0

Jan Hajič, Marie Mikulová,
Jaroslava Hlaváčová, Milan Straka, Eduard Bejček,
Jan Štěpánek
et al.

LDC 2020

text **PDT**
PDTSC speech
translation **PCEDT**
FAUST internet
Morphology
Syntax
Semantics

ÚFAL

LINDAT CLARIN

Prague Dependency Treebank Consolidated PDT-C 1.0

Jan Hajič, Marie Mikulová,
Jaroslava Hlaváčová, Milan Straka,
Jan Štěpánek, Barbora Štěpánková
et al.

LDC 3020

text **PDT**
PDTSC speech
translation **PCEDT**
FAUST internet
Morphology
Syntax
Semantics

ÚFAL

LINDAT CLARIN

PML-TQ

Matyáš Kopp
kopp@ufal.mff.cuni.cz

What is PML-TQ

Relations ▾ Node Types ▾ Attributes ▾ Operators ▾ Functions ▾

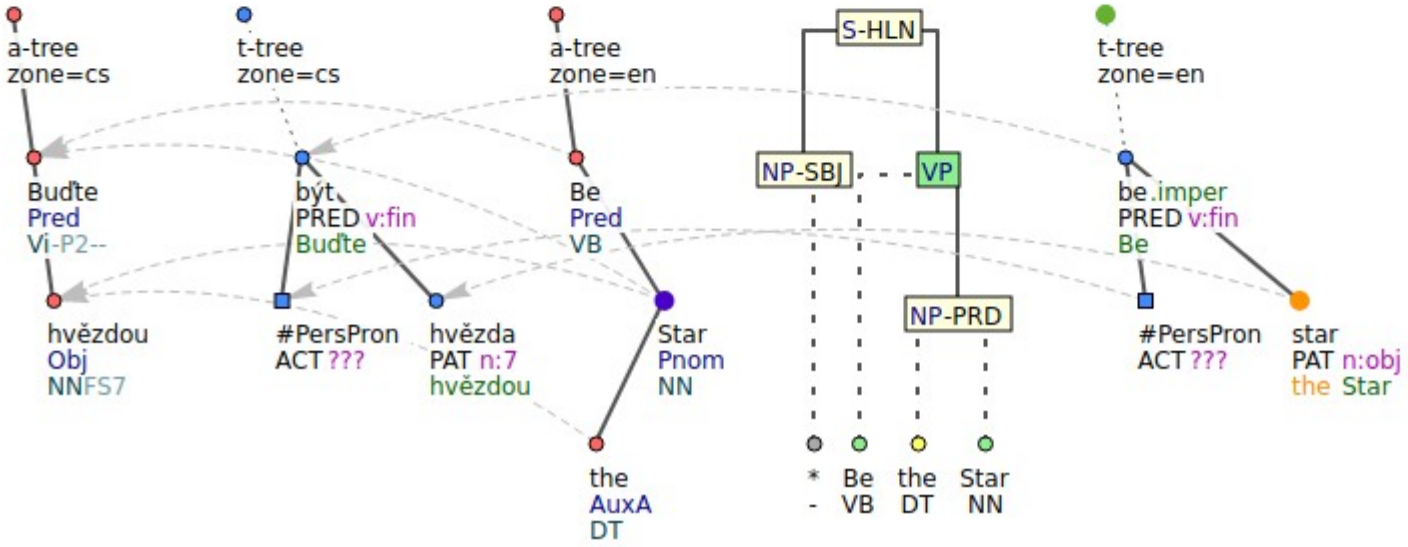
```
t-root $a := [
  same-tree-as t-node $c := [
    a/lex.rf a-node $d:=[
      ord = 3,
      ! same-tree-as a-node $e :=[
        ord > $d.ord
      ]
    ],
  ],
];
```

Execute query | w/o Filters | Suggest (0)

Result: ← 51 / 100 →

1 t-root $a | 2 t-node $c | 3 a-node $d

[en] Be the Star
[cs] Buďte hvězdou



PML-TQ Clients

- Web (available at https://lindat.mff.cuni.cz/services/pmltq/)
- TrEd extension
- Command line interface (part of perl PMLTQ module)
  - `pmltq query --btred --query 'a-node [] >> count()' *.a.*`
    - btred should be in PATH
  - `pmltq query --server a --query 'a-node [] >> count()'`
    - ‘a’ is treebank name in TrEd configuration file.
      It can be replaced link:

      `http://euler.ms.mff.cuni.cz/api/treebanks/pdt30`

Plan

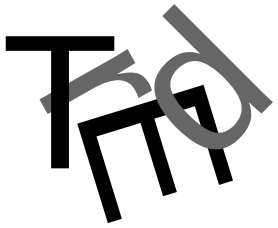
- Web Interface
  - Tree visualization
    - Node inspection
    - Zoom
    - Better matched node highlight
  - Language editing
  - Query list caching
- CLI
  - Unify and simplify options
- PMLTQ core Perl module
  - Asynchronuous SQL evaluation
  - Split to multiple CPAN packages
- Documentation
- PML-TQ for Linguistic Data Consortium

In case of problems

- kopp@ufal.mff.cuni.cz
- pmltq@ufal.mff.cuni.cz
- https://github.com/ufal/perl-pmltq-web/issues
- https://github.com/ufal/perl-pmltq/issues
- https://github.com/ufal/perl-pmltq-server/issues

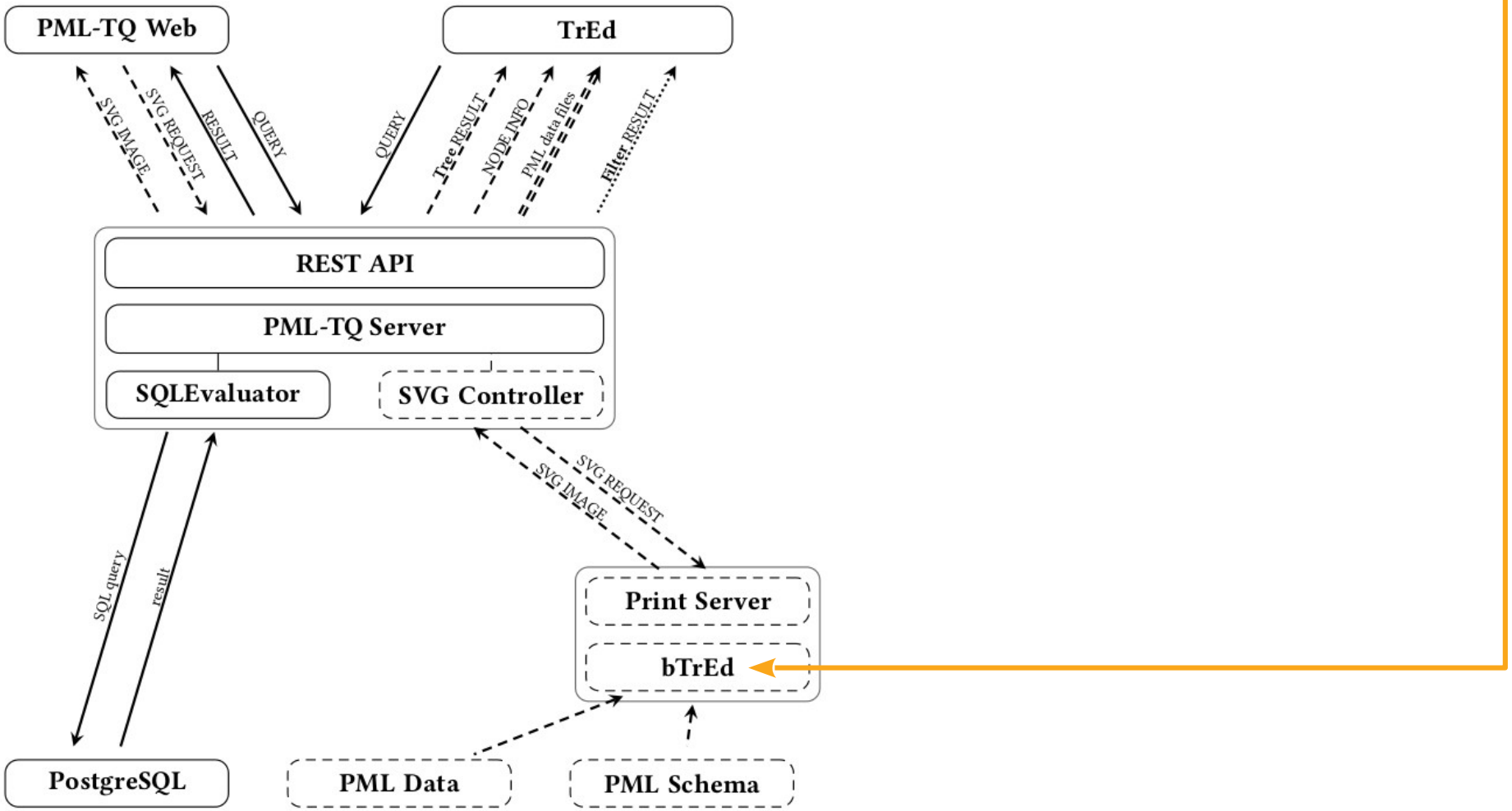
TrEd

Matyáš Kopp
kopp@ufal.mff.cuni.cz



Why me?

- PML-TQ tree visualization depends on TrEd



Current state of ‘development’

- It is possible to install TrEd on:
  - Windows (tested on win10, no signature)
  - Linux (perlbrew — problem with architecture dependent perl packages)
  - Mac OS (no signature)
- Minor changes needed by PML-TQ (tree visualization)
- Plan:
  - Repository cleaning
  - Adding signature

In case of problems

- Releasing, installation, runtime errors:
  - kopp@ufal.mff.cuni.cz
- Extension development
  - devel@ufal.mff.cuni.cz — there are more experienced TrEd users and extension developers than me

kontext

Query Corpora Save Concordance Filter Frequency Collocations View

Corpus: Prague Dependency Treebank 3.5 | **Query**: .*vyhled.* (30 hits)

Hits: **30** | i.p.m.: **36.01** (related to the whole "pdt_35_cs_t") | ARF: **18.46** | Result is sorted 1 / 1

Line selection: simple ▾ | Attribute viewing mode:

	v polovině dubna v Praze	**vyhledala**	Božena Dvořáková , zaměstnankyně ústředního
	v zahraničí . Jsou to	**nejvyhledávanější**	služby komory , řekl .
	malým a středním podnikům při	**vyhledávání**	zahraničního partnera ve 43 zemích
	v poslední době zaměřil na	**vyhledávání**	hluchoslepých . Nedokážeme odhadnout ,
	než si dát práci a	**vyhledat**	z bank tu nejlepší a
	své knize se především snažím	**vyhledat**	styčné body mezi šerem tajemství
	transsexuálů , kteří sexuologickou ordinaci	**nevyhledají**	a sami se smíří s
	v diáři i opravovat nebo	**vyhledávat**	a součástí programu je i
	účtu , dispozičním právem a	**vyhledávání**	v archívu . Zablokování účtu
	Zájemci mohou požádat i o	**vyhledání**	vhodného partnera v zahraničí na

kontext

Collocations View Hel

Corpus: Prague Depende

Hits: 30 | i.p.m.: 36.01 (

1 / 1

Line selection: simple

[PDT Vallex 📖]

vyhledat ACT(1) PAT(4)
najít, nalézt

- vyhledat lékařské ošetření

Show examples

[CzEngVallex 📖]

vyhledat : seek
vyhledat ACT(1) PAT(4)
najít, nalézt

- vyhledat lékařské ošetření

Show details for pair 1

vyhledat : seek_out
vyhledat ACT(1) PAT(4)
najít, nalézt

- vyhledat lékařské ošetření

Show details for pair 1

aměstnankyně ústředního

ve 43 zemích

kážeme odhadnout ,

m tajemství

trans

e i

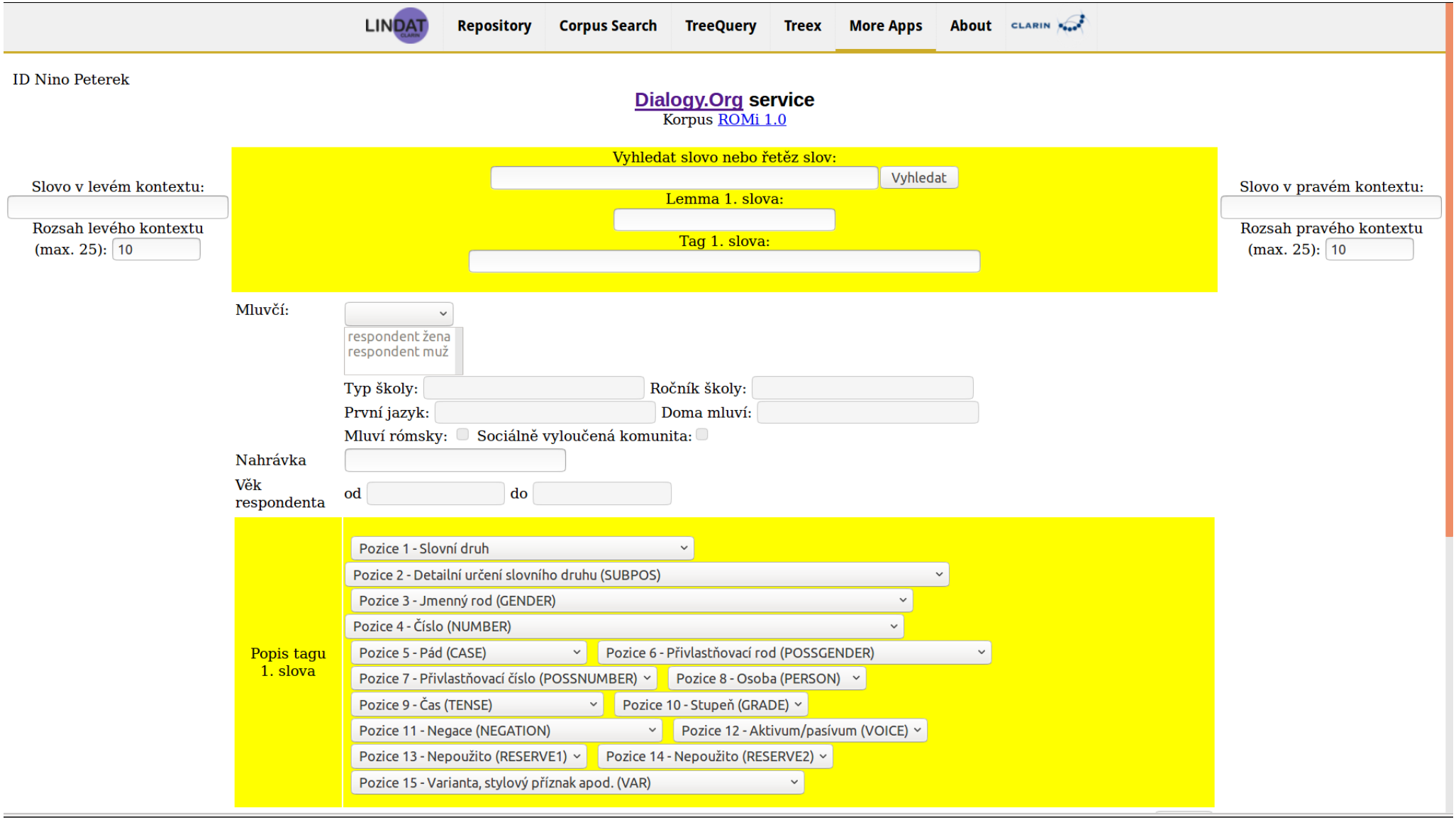
í účtu

Zájemci mohou požádat o vyhledání vhodného partnera v zahraničí na

Summa teologie	Summa Theologiae
Při první se postupuje takto : Zdá se , že není třeba , aby mimo filosofické předměty byla ještě nauka jiná . Člověk totiž nemá usilovati o to , co jest nad rozum , podle onoho Eccli . 22 : " **Nevyhledávej** , co jest nad tebe " . Ale filosofické nauky dostatečně pojednávají o tom , co jest podřízeno rozumu . Zdá se tedy zbytečné , míti jinou nauku mimo filosofické předměty .	Ad primum sic proceditur . Videtur quod non sit necessarium , praeter philosophicas disciplinas , aliam doctrinam haberi . Ad ea enim quae supra rationem sunt , homo non debet conari , secundum illud Eccli. III , altiora te ne quaesieris . Sed ea quae rationi subduntur , sufficienter traduntur in philosophicis disciplinis . Superfluum igitur videtur , praeter philosophicas disciplinas , aliam doctrinam haberi .
K prvnímu se tedy musí říci , že , ačkoli člověk nemá rozumem **vyhledávati** , co jest nad lidské poznání , musí to přece přijmouti věrou , když bylo Bohem zjeveno . Proto se tamtéž pokračuje : " Velmi mnoho nad mysl člověka ti bylo ukázáno " . A v tom záleží posvátná nauka .	Ad primum ergo dicendum quod , licet ea quae sunt altiora hominis cognitione , non sint ab homine per rationem inquirenda , sunt tamen , a Deo revelata , suscipienda per fidem . Unde et ibidem subditur , plurima supra sensum hominum ostensa sunt tibi . Et in huiusmodi sacra doctrina consistit .
Mimo to Filosof praví na počátku Metaf . , že badavé vědy se **vyhledávají** pro ně samé . A přece se nemůže	Praeterea , philosophus dicit , in principio Metaphys. , quod scientiae speculativae propter seipsas quaeruntur . Nec

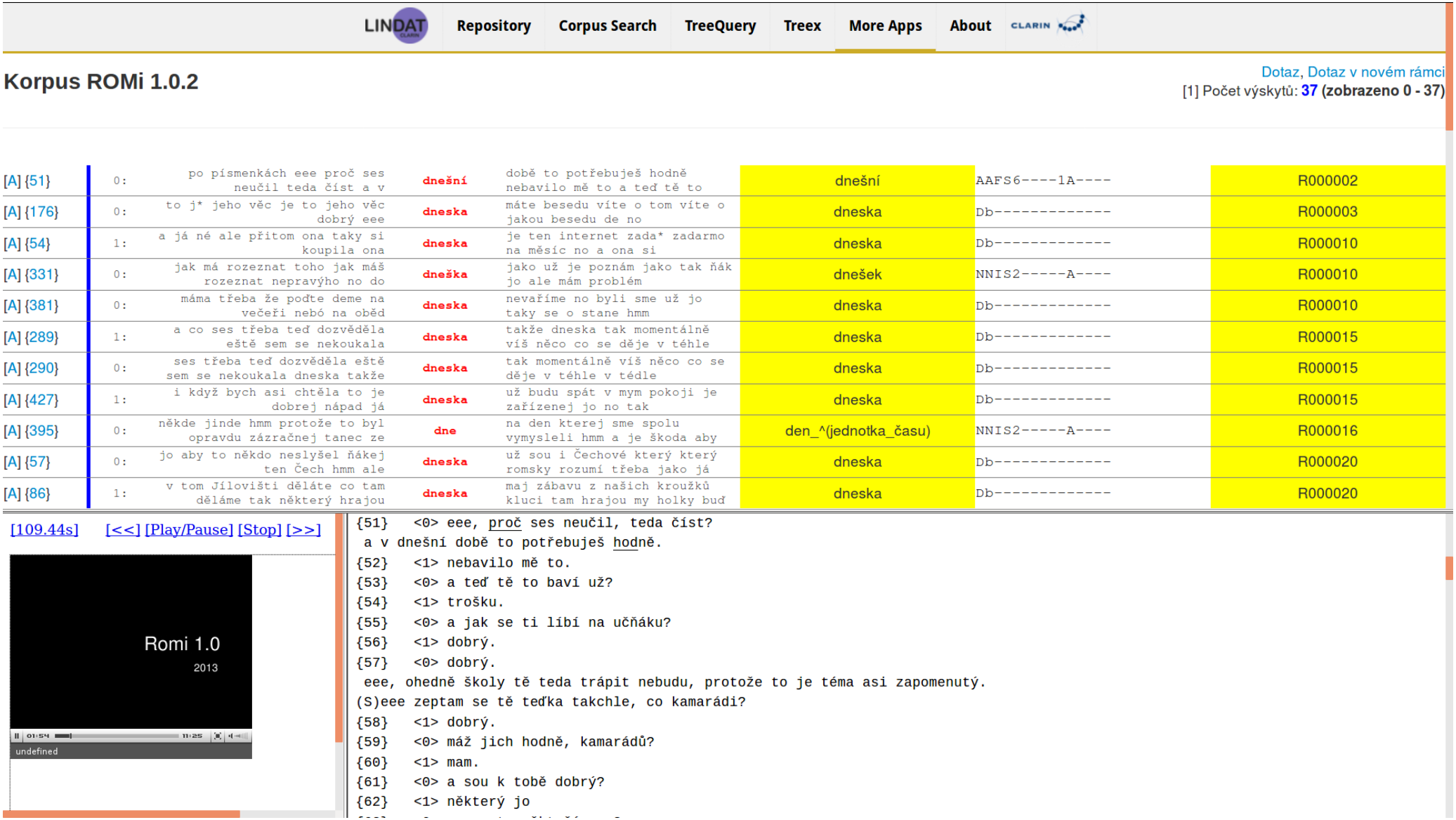
Dialogy.Org
search engine

Korpus ROMi 1.0
http://lindat.mff.cuni.cz/services/dialogy.org



Dialogy.Org
search engine

Korpus ROMi 1.0
http://lindat.mff.cuni.cz/services/dialogy.org



Dialogy.Org
search engine

Korpus DIALOG 1.2
http://ujc.dialogy.cz

Korpus Monolog 1.1
http://monolog.dialogy.org

[V1] [F0] {33}	Václav Klaus:	země vůči evropské unii mně se zdá že v tomto	dialogu	který mimochodem nonstop probíhá v českém parlamentu a zúčastňují se ho
[V1] [F0] {257}	Lubomír Zaorálek:	práce je v obrovské defenzívě povinností naší dneska je sociální	dialog	a usilování o sociální smír ano ale ten právě to co
[V1] [F0] {264}	Jan Kasal:	poslední větě kterou ste řekl že velmi usilujete o sociální	dialog	přece sociální demokracie tento sociální dialog vůbec nevedla vy chcete tvrdit
[V1] [F0] {264}	Jan Kasal:	velmi usilujete o sociální dialog přece sociální demokracie tento sociální	dialog	vůbec nevedla vy chcete tvrdit že počkejte pane předsedo že že
[V1] [F0] {267}	Jan Kasal:	pane předsedo že že tento zákoník práce nevzešel na základě	dialogu	mezi sociálními partnery v tom mně dáte určitě za pravdu a
[V1] [F0] {267}	Jan Kasal:	rozhodli pro převálcování ve prospěch jedné skupiny partnerů toho sociálního	dialogu	to nám je hrozně líto protože tím válcovacím mechanismem ste dali
[V1] [F0] {282}	Bohumil Klepl:	mm s prominutím nasrávací typ takže já se dokážu v	dialogu	sám se sebou tak namíchnout (že normálně ((i křičím
[V1] [F0] {89}	Magdalena Kožená:	byl takový rečitativ a tak sme si vyměňovali nějaký ten	dialog	a já sem vlastně nesměla ani hýbat pusou protože režisérovi se
[V1] [F0] {91}	Karol Sidon:	možná nedocenil prostě jako výsledky třeba toho židovsko nebo křesťansko-židovskýho	dialogu	kterýmu sem se v podstatě jako teologicky vyhýbal poněvadž sem to
[V1] [F0] {95}	Karol Sidon:	to si myslim dál že vlastně jako třeba tenhleten způsob	dialogu	teologickýho který mě osobně ňák zvlášť ne nezajímá mohl obrousit prost
[V1] [F0] {119}	Miloslav Vlk:	učení jana pavla druhého kde základním slovem je otevřenost nebo	dialog	třeba jo a tato společnost ve které žijeme mně nutí abych

[1137.1s] [<<] [Play/Pause] [Stop] [>>]

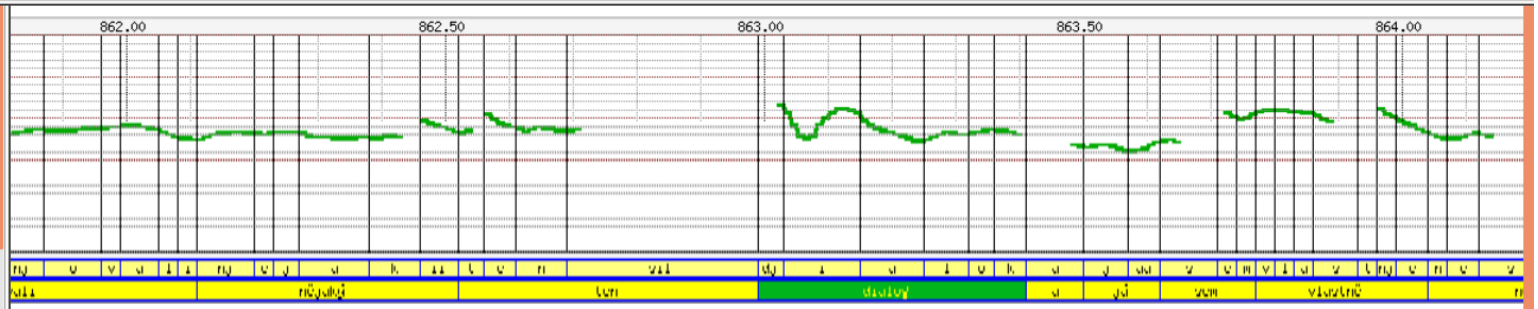
{282} BK: nahlas, takže kdyby mě někdo sledoval zvenčí tak si myslí že tam
někoho mám, a já tam nikoho nemám, a já sem takovej mm ee (.) s prominutím
nasrávací typ takže já se dokážu v dialogu sám se sebou tak namíchnout, (že
normálně (

{283} EK: (((smích))(

{284} MM: [((smích))]

{285} BK: i křičím, a říkám ty di do hajzlu hele s tebou se nebudu bavit, a
pak dělám to že když vypiju třeba: i když teď už nemůžu protože mám pocit že
sem trošku nemocnej na ty vnitřnosti, tak ee ee začnu e mobilním telefonem e
psát takové litostivé ruské esemesky,

[861.99s] [<<] [Play/Pause] [Stop] [>>]



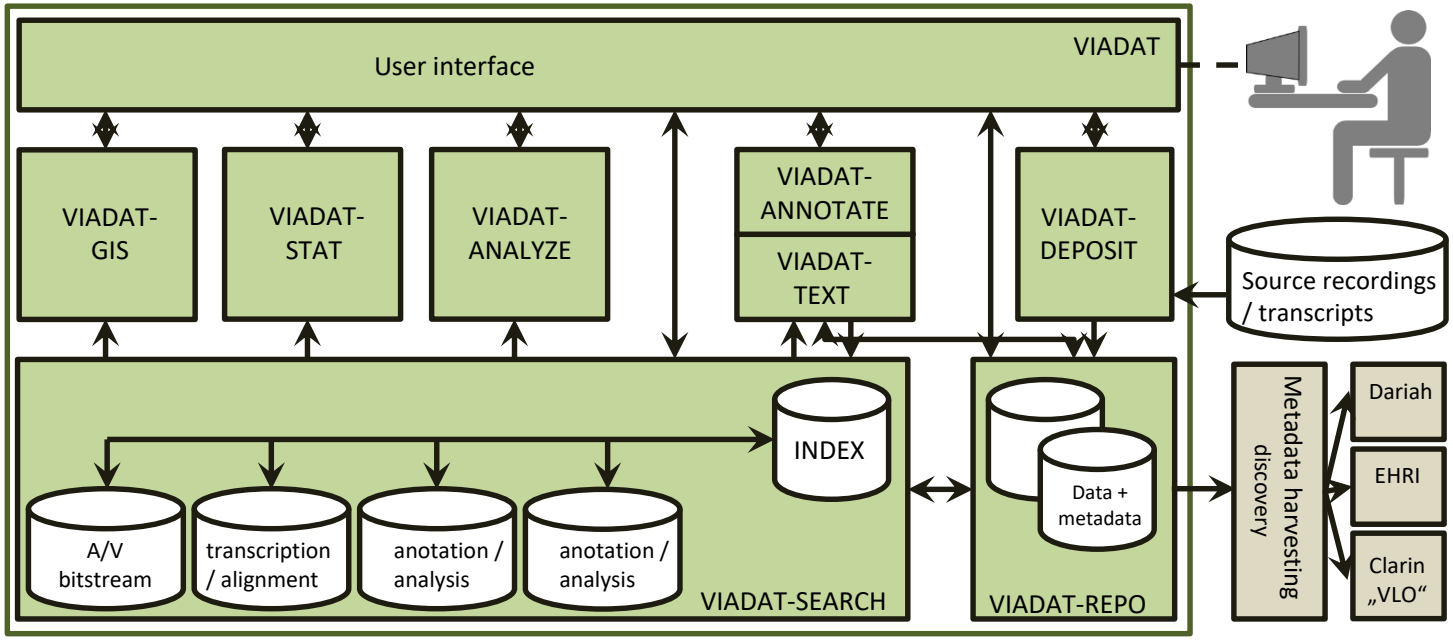
Transferring data from ujc.dialogy.cz...

NAKI II:
“VIADAT” and “USTR” projects

Jan Hajič

VIADAT (2016-2019)

- Ministry of Culture Applied projects program
- VIADAT: Virtual Assistant for Oral Archives
  - With National Film Archive, Institute of Comtemp. History



“USTR” project

- Same program, project led by West Boh. Univ.
  - With USTR – Institute for the Study of Totalitarian Regimes
- UFAL/LINDAT: Supplying tools only
- Goal: search in audio/video recordings
  - Similar to the MALACH project / CVHM
- Also 2016-2019

CEMI - Center for Large Scale Multi-modal Data Interpretation

Funding: Czech Science Foundation (GAČR),14 Million CZK

Duration: 7 Years, 2012 - 2018

Consortium: 4 Institutions

1. ČVUT: Czech Technical University, Jiří Matas (coordinator)
2. MU: Masaryk University, Pavel Zezula
3. ZČU: University of West Bohemia, Josef Psutka
4: UK: Charles University, Pavel Pecina

Projects goals

The project aims at exploiting **large** collections of **unlabeled multi-modal data**, mainly video footage, to further state-of-the-art in **video**, **audio** and **natural language** understanding, interpretation, annotation and retrieval by combining **unsupervised** and **semi-supervised** learning.

Expertise:

- ČVUT: image processing (video, pictures)
- ZČU: speech processing (speech recognition)
- MU: similarity search (images, text)
- UK: NLP, MT …

UFAL contribution/outcomes

Staff

- Silvie Cinková, **Jan Hajič jr. (2018)**, **Petra Galuščáková (2017)**, Jindřich Helcl, Martin Holub, Ema Krejčová, **Jindřich Libovický (2018)**, Tomáš Musil, Pavel Pecina, Lenka Smejkalová, Anna Vernerová

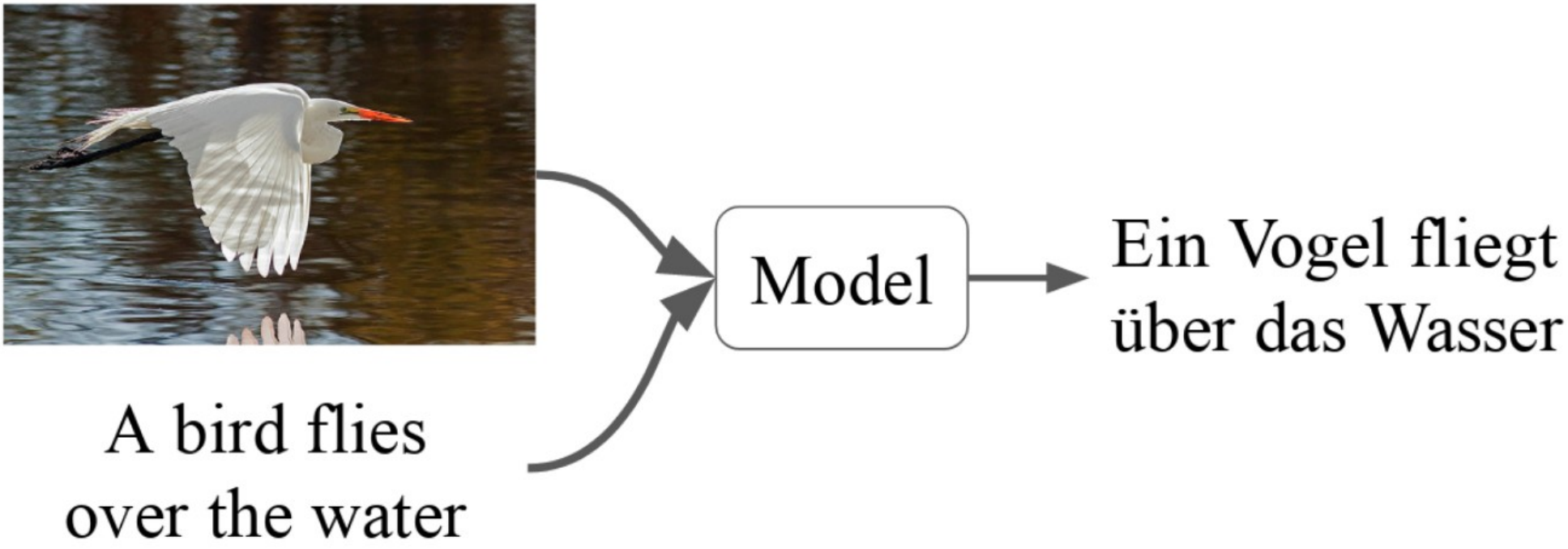
Publications

- ~72 items in Biblio
- 5 IF journals
- 2x A*, 6x A, 5x B, 5x C conference papers (CORE ranking)

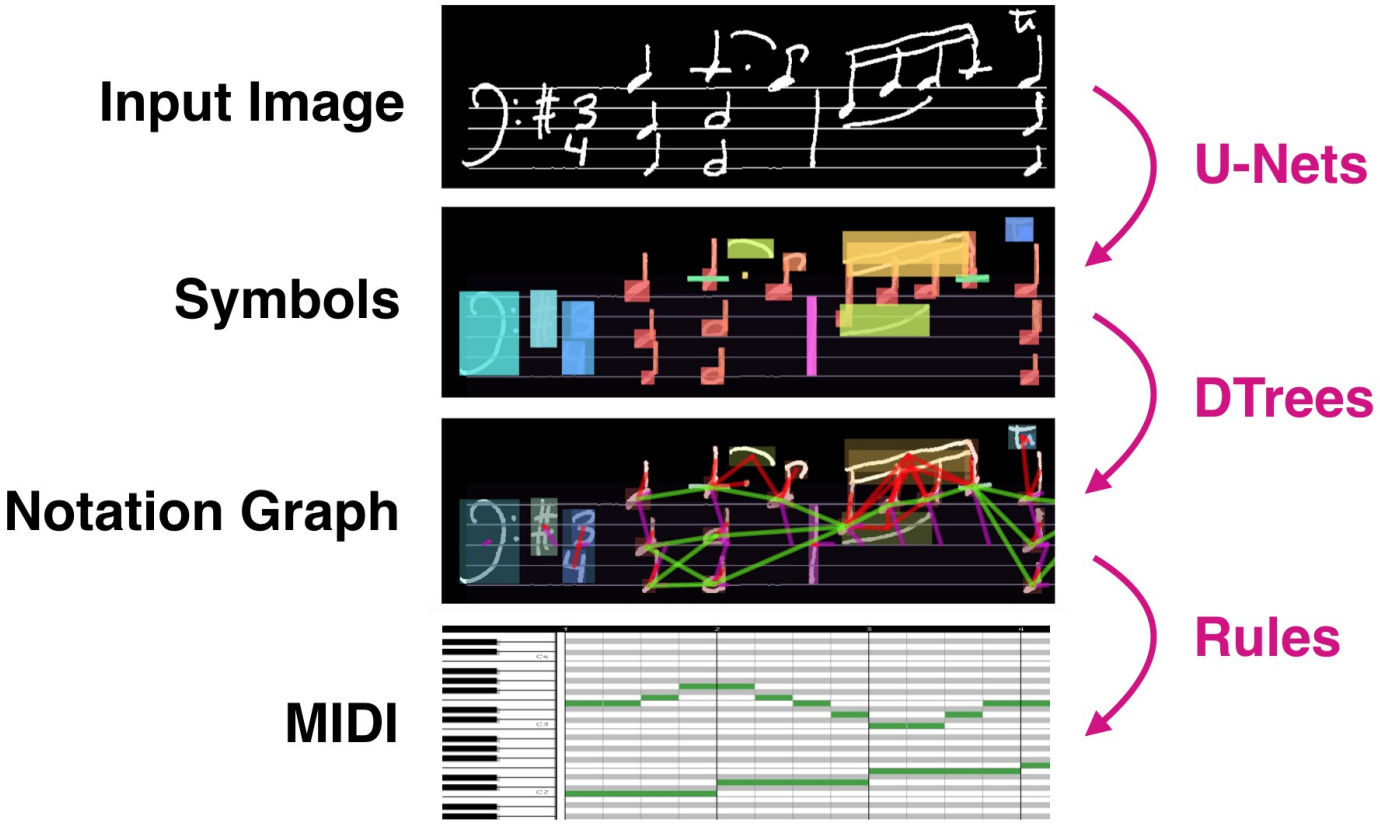
- **Tasks**
  - text recognition from scene images, speech retrieval, search/hyperlinking in audiovisual data, natural language identification (text, speech), cross-lingual information retrieval

Scene text recognition
(J. Libovický)

Multimodal machine translation
(J. Libovický, J. Helcl)



Handwritten Music Recognition
(J. Hajič jr.)



Seach and hyperlinking in audio-video
(P. Galusčáková)

CEMI TechDemo

Korektor

Spellchecker and an occasional grammarchecker

- provides spellchecking with automatic context-aware corrections

Korektor

Spellchecker and an occasional grammarchecker

- provides spellchecking with automatic context-aware corrections
- diacritics restoration most reliable

Korektor

Spellchecker and an occasional grammarchecker

- provides spellchecking with automatic context-aware corrections
- diacritics restoration most reliable
- general spellchecking module capable of correcting spelling errors and (ideally) also grammatical errors

Korektor

Spellchecker and an occasional grammarchecker

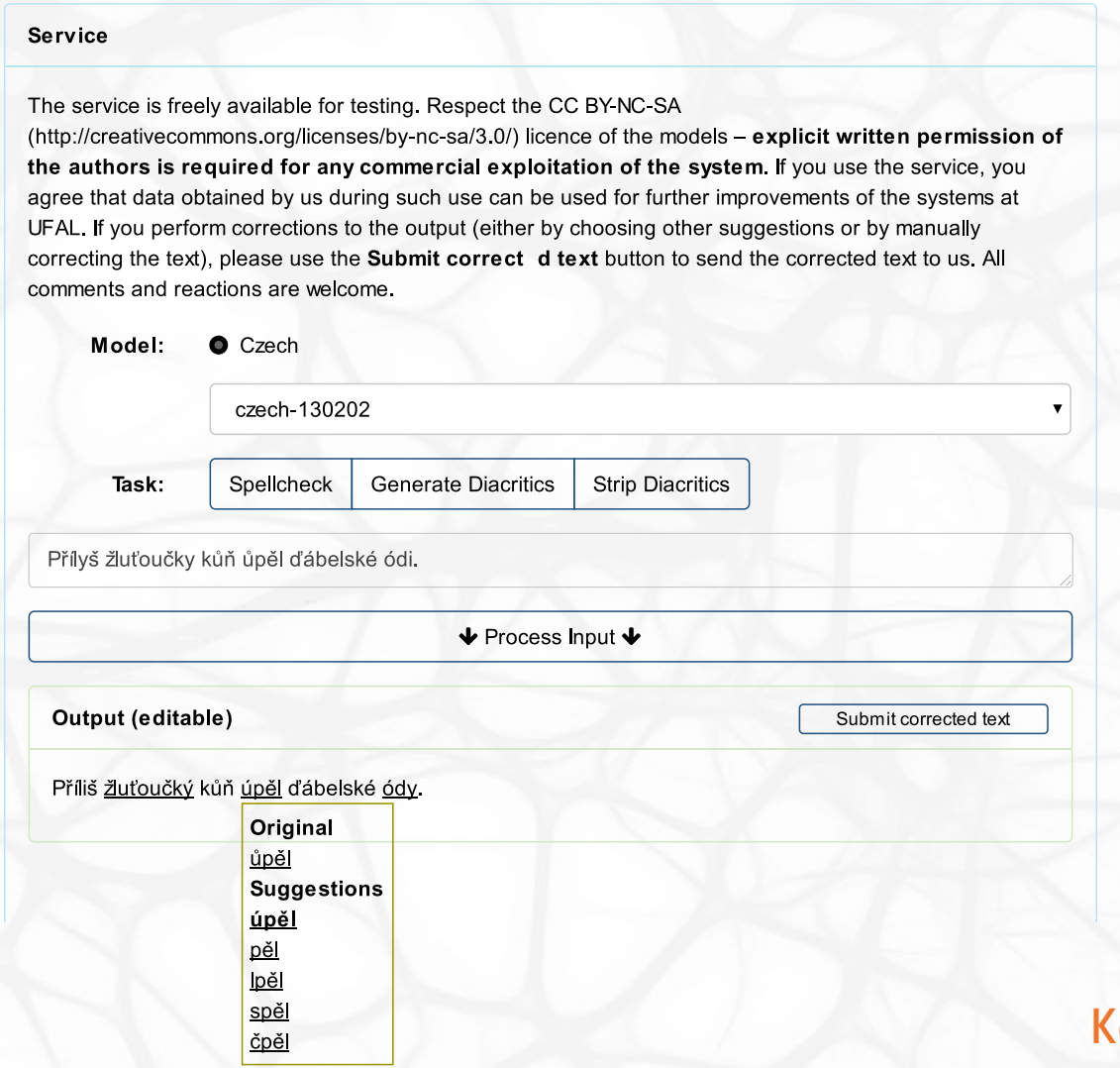
- provides spellchecking with automatic context-aware corrections
- diacritics restoration most reliable
- general spellchecking module capable of correcting spelling errors and (ideally) also grammatical errors

Available as

- binary

Korektor

- web REST service



Korektor

- browser plugin allowing to correct any editable field

Korektor

- the current version is based on Michal Richter's Master thesis (2012)

Korektor

- the current version is based on Michal Richter's Master thesis (2012)
- models primarily for Czech, difficult to extend for other languages
  - explicit morphological dictionary is needed
  - explicit error model definition is required

Korektor

- the current version is based on Michal Richter's Master thesis (2012)
- models primarily for Czech, difficult to extend for other languages
  - explicit morphological dictionary is needed
  - explicit error model definition is required

Future Plans

- Jakub Náplava is developing a new version based on deep NNs:
  - better performance
  - easily extensible

MorphoDiTa

Morphological **di**ctionary and **ta**gger

MorphoDiTa

Morphological **di**ctionary and **ta**gger

Morphological dictionary is a set of triples

(form, lemma, tag)

Morphological **di**ctionary and **ta**gger

Morphological dictionary is a set of triples

(form, lemma, tag)

- efficiently encodes it (2MB out of 6.5GB)

MorphoDiTa

Morphological **di**ctionary and **ta**gger

Morphological dictionary is a set of triples

(form, lemma, tag)

- efficiently encodes it (2MB out of 6.5GB)
- allows performant lookups
  - morphological analysis (all matching triples for a given form)
  - morphological generation (all matching triples for a given lemma)

Morphological dictionary and tagger

Morphological dictionary is a set of triples

(form, lemma, tag)

- efficiently encodes it (2MB out of 6.5GB)
- allows performant lookups
  - morphological analysis (all matching triples for a given form)
  - morphological generation (all matching triples for a given lemma)
- provides a *guesser* (currently only for morphological analysis)

MorphoDiTa

Can be combined with a derivational dictionary (e.g., *DeriNet*) in a form of tuples of

(lemma, derivational parent lemma)

MorphoDiTa

Can be combined with a derivational dictionary (e.g., *DeriNet*) in a form of tuples of

(lemma, derivational parent lemma)

- provides traversal in derivational tree
  - parent
  - children

MorphoDiTa

Can be combined with a derivational dictionary (e.g., *DeriNet*) in a form of tuples of

(lemma, derivational parent lemma)

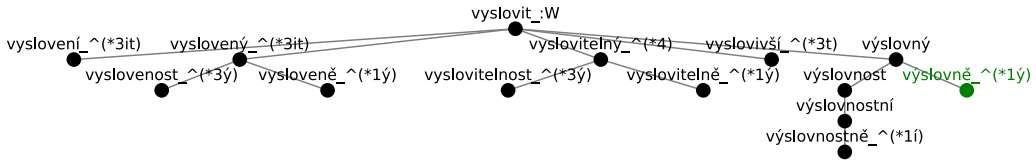
- provides traversal in derivational tree
  - parent
  - children
- furthermore, lemmas returned by MorphoDiTa might be replaced by
  - derivational roots
  - all lemmas on path to derivational roots
  - whole derivational trees

MorphoDiTa

Can be combined with a derivational dictionary (e.g., *DeriNet*) in a form of tuples of

(lemma, derivational parent lemma)

- provides traversal in derivational tree
  - parent
  - children
- furthermore, lemmas returned by MorphoDiTa might be replaced by
  - derivational roots
  - all lemmas on path to derivational roots
  - whole derivational trees

Token	Lemma	Tag	Derivation
výslovně	výslovně_^(*1ý)	Dg-------1A----	

MorphoDiTa

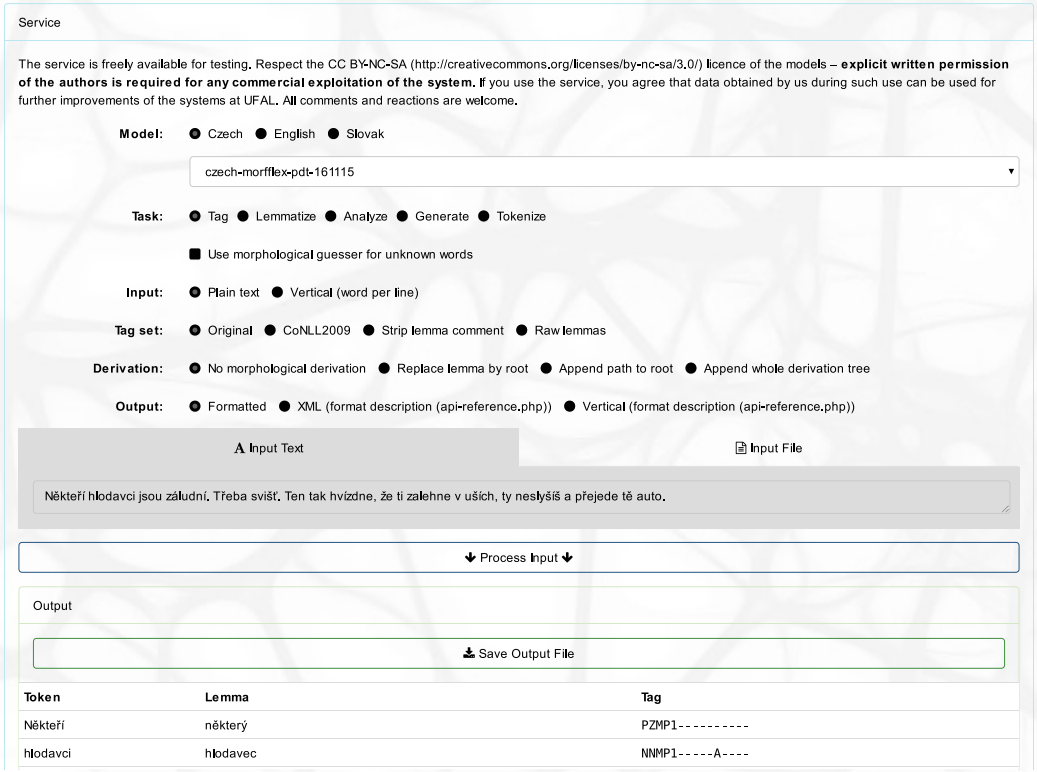
The tagging part can perform disambiguation of POS tags and lemmas.

MorphoDiTa

The tagging part can perform disambiguation of POS tags and lemmas.

Technical stuff

- efficient C++ implementation under open-source license (MPL 2.0)
- bindings for Python, Perl, Java, C#
- web REST service



MorphoDiTa

Future Plans

- improved tagging models (results in another slot about deep learning)

Future Plans

- improved tagging models (results in another slot about deep learning)
- better guesser for morphological analysis

Future Plans

- improved tagging models (results in another slot about deep learning)
- better guesser for morphological analysis
- a guesser for morphological generation

Future Plans

- improved tagging models (results in another slot about deep learning)
- better guesser for morphological analysis
- a guesser for morphological generation
- move to CoNLL-U as the universal format

Future Plans

- improved tagging models (results in another slot about deep learning)
- better guesser for morphological analysis
- a guesser for morphological generation
- move to CoNLL-U as the universal format
- provide binary wheels for Python

UDPipe

Completely trainable pipeline for analysing texts

UDPipe

Completely trainable pipeline for analysing texts

- uses CoNLL-U as internal representation

UDPipe

Completely trainable pipeline for analysing texts

- uses CoNLL-U as internal representation
- performs
  - tokenization
  - sentence segmentation
  - POS tagging into UPOS, XPOS and FEATS
  - lemmatization
  - dependency parsing

UDPipe

Completely trainable pipeline for analysing texts

- uses CoNLL-U as internal representation
- performs
  - tokenization
  - sentence segmentation
  - POS tagging into UPOS, XPOS and FEATS
  - lemmatization
  - dependency parsing
- everything trainable from data only

UDPipe

Completely trainable pipeline for analysing texts

- uses CoNLL-U as internal representation
- performs
  - tokenization
  - sentence segmentation
  - POS tagging into UPOS, XPOS and FEATS
  - lemmatization
  - dependency parsing
- everything trainable from data only
- pretrained models for UD data (60 languages, 90 treebanks in UD 2.2)

UDPipe

Completely trainable pipeline for analysing texts

- uses CoNLL-U as internal representation
- performs
  - tokenization
  - sentence segmentation
  - POS tagging into UPOS, XPOS and FEATS
  - lemmatization
  - dependency parsing
- everything trainable from data only
- pretrained models for UD data (60 languages, 90 treebanks in UD 2.2)
- technical stuff
  - efficient C++ implementation under open-source license (MPL 2.0)
  - bindings for Python, Perl, Java, C#
  - web REST service

UDPipe

Usage

- used in CoNLL 2017 and CoNLL 2018 Shared Tasks in UD parsing

UDPipe

Usage

- used in CoNLL 2017 and CoNLL 2018 Shared Tasks in UD parsing
- current UDPipe 2.0 prototype
  - one of three winners of CoNLL 2018 ST

UDPipe

Usage

- used in CoNLL 2017 and CoNLL 2018 Shared Tasks in UD parsing
- current UDPipe 2.0 prototype
  - one of three winners of CoNLL 2018 ST
  - overall winner of Extrinsic parser evaluation, EPE 2018

UDPipe

Usage

- used in CoNLL 2017 and CoNLL 2018 Shared Tasks in UD parsing
- current UDPipe 2.0 prototype
  - one of three winners of CoNLL 2018 ST
  - overall winner of Extrinsic parser evaluation, EPE 2018
  - more detailed results in another slot about deep learning

UDPipe

Usage

- used in CoNLL 2017 and CoNLL 2018 Shared Tasks in UD parsing
- current UDPipe 2.0 prototype
  - one of three winners of CoNLL 2018 ST
  - overall winner of Extrinsic parser evaluation, EPE 2018
  - more detailed results in another slot about deep learning
- the web REST service processed ~750 000 requests in last 5 months, processing ~3GB of data

Future Plans

- release improved models
  - GPU for inference is recommended

Future Plans

- release improved models
  - GPU for inference is recommended
- improve tokenization
  - allow documents to contain formatting markup, e.g., <b> etc.

Future Plans

- release improved models
  - GPU for inference is recommended
- improve tokenization
  - allow documents to contain formatting markup, e.g., <b> etc.
  - allow processing only some parts of documents, e.g., only content of <text> elements etc.
- use CoNLL-U as the universal format in our services, with extensions in form of sentence-level and document-level object data
  - bundle with appropriate tools like NER, sentiment analysis, …

Future Plans

- release improved models
  - GPU for inference is recommended
- improve tokenization
  - allow documents to contain formatting markup, e.g., <b> etc.
  - allow processing only some parts of documents, e.g., only content of <text> elements etc.
- use CoNLL-U as the universal format in our services, with extensions in form of sentence-level and document-level object data
  - bundle with appropriate tools like NER, sentiment analysis, …
- start with more difficult (more semantic) tasks
  - coreference resolution (à la English)

Future Plans

- release improved models
  - GPU for inference is recommended
- improve tokenization
  - allow documents to contain formatting markup, e.g., <b> etc.
  - allow processing only some parts of documents, e.g., only content of <text> elements etc.
- use CoNLL-U as the universal format in our services, with extensions in form of sentence-level and document-level object data
  - bundle with appropriate tools like NER, sentiment analysis, …
- start with more difficult (more semantic) tasks
  - coreference resolution (à la English)
  - t-layer parsing

Cross-lingual Syntactic Parsing

Cross-lingual Syntactic Parsing

- e.g.: parse Lower Sorbian

Cross-lingual Syntactic Parsing

- e.g.: parse Lower Sorbian
  - no training data

Cross-lingual Syntactic Parsing

- e.g.: parse Lower Sorbian
  - no training data
  - use training data for other close languages
    - Upper Sorbian, Polish, Czech

Cross-lingual Syntactic Parsing

- e.g.: parse Lower Sorbian
  - no training data
  - use training data for other close languages
    - Upper Sorbian, Polish, Czech
- our focus: dependency parsing, POS tagging
- probably adaptable to other NLP tasks
  - no/low annotated data for the “target” language
  - annotated data for some (close) “source” languages

Methods, subtasks, problems

Methods, subtasks, problems

- strong multilingualism
  - massively multiparallel corpora (Bible, Watchtower)
  - tens or hundreds of languages

Methods, subtasks, problems

- strong multilingualism
  - massively multiparallel corpora (Bible, Watchtower)
  - tens or hundreds of languages
- word alignment, machine translation
  - low data
  - special needs (very literal, monotone, 1:1…)

Methods, subtasks, problems

- strong multilingualism
  - massively multiparallel corpora (Bible, Watchtower)
  - tens or hundreds of languages
- word alignment, machine translation
  - low data
  - special needs (very literal, monotone, 1:1…)
- cross-lingually consistent annotation (UD)

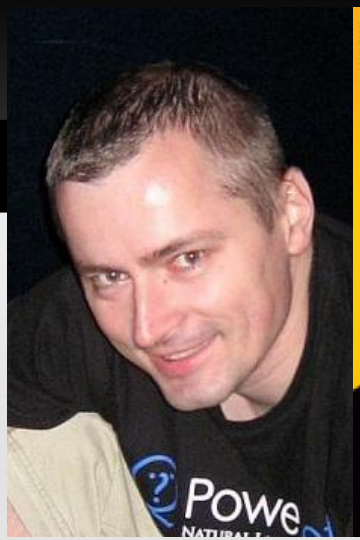
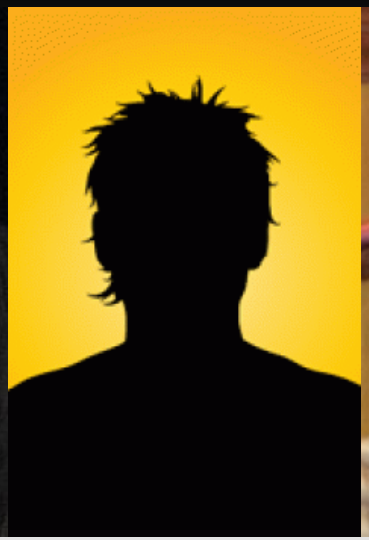
Methods, subtasks, problems

- strong multilingualism
  - massively multiparallel corpora (Bible, Watchtower)
  - tens or hundreds of languages
- word alignment, machine translation
  - low data
  - special needs (very literal, monotone, 1:1…)
- cross-lingually consistent annotation (UD)
- tokenization, encoding, transliteration…

At ÚFAL

- The “team”
  - Dan Zeman
  - Zdeněk Žabokrtský, David Mareček, Rudolf Rosa
  - Loganathan, Adedayo Oluokun, Vinit Ravishankar…

At ÚFAL

- The “team”
  - Dan Zeman
  - Zdeněk Žabokrtský, David Mareček, Rudolf Rosa
  - Loganathan, Adedayo Oluokun, Vinit Ravishankar…
- Best in the world (at least in shared tasks)
  - 2017 VarDial: 1st place
  - 2018 CoNLL: 1st place (for low-resource languages)

At ÚFAL

- The “team”
  - Dan Zeman
  - Zdeněk Žabokrtský, David Mareček, Rudolf Rosa
  - Loganathan, Adedayo Oluokun, Vinit Ravishankar…
- Best in the world (at least in shared tasks)
  - 2017 VarDial: 1st place
  - 2018 CoNLL: 1st place (for low-resource languages)
- NPFL120 Multilingual NLP (Dan, Rudolf, oBo)
  - summer semester, everyone welcome!

NER and NameTag

Jana Straková
strakova@ufal.mff.cuni.cz

Kostelec nad Ohří, 2018

NameTag currently

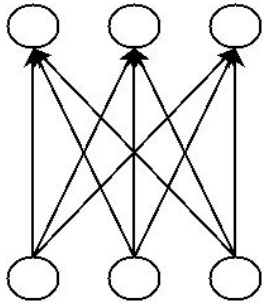
Named Entity Recognition tool,

open-source, free software (Mozilla Public License 2.0),

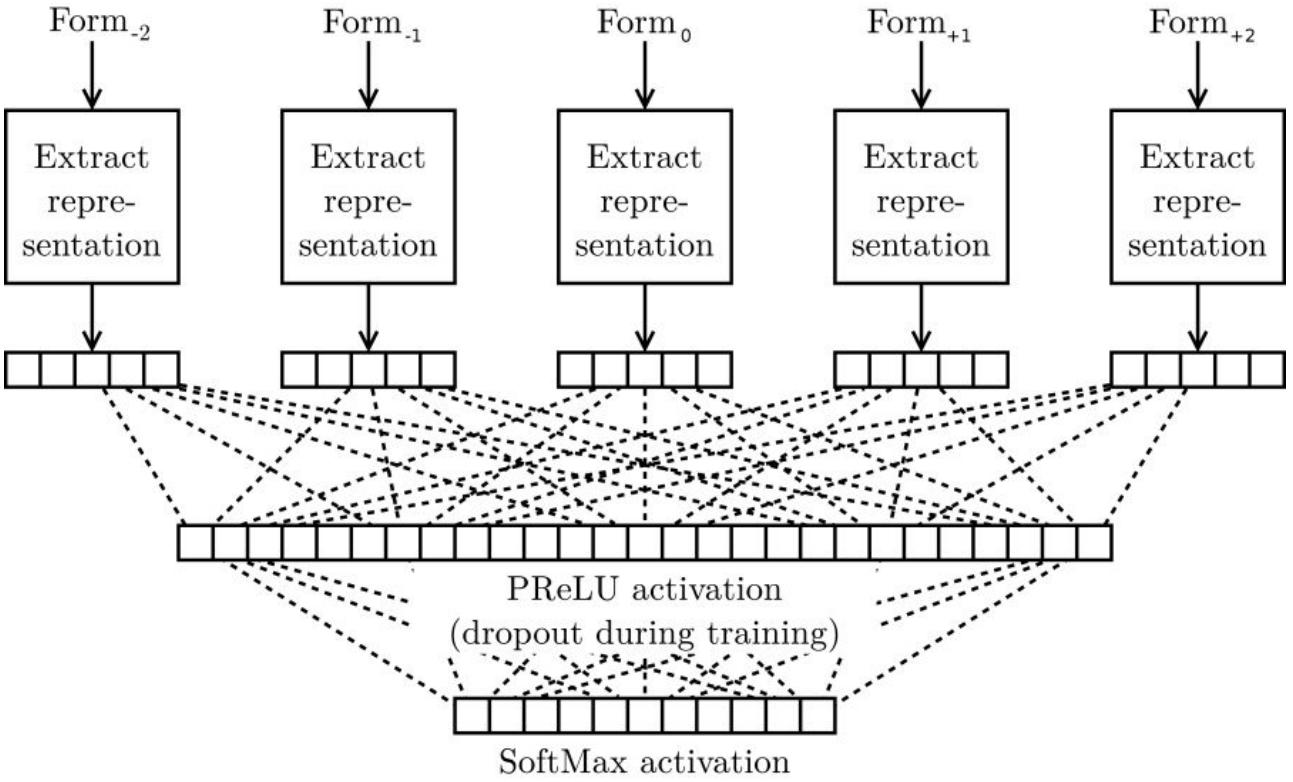
precompiled binaries, pretrained models for Czech and English (CC BY-NC-SA),

REST web service,

bindings in Python, Perl.



NER Thesis 2017



NameTag without morphology

Improved architecture with deep NN bells and whistles:

(Deep NN, lazy Adam, RNNs with LSTM, pretrained WEs, end-to-end WEs, end-to-end CLEs with GRU, dropout, word dropout, adaptive learning rate.)

No manual classification features.

No morphological analysis, no POS, just forms.

	Thesis 2017 (Straková et. al, 2016)	09/2018
CNEC 1.0 F1	73.71	78.70

NameTag (Near :)) Future

Release new architecture (trade-off between F1 and model size/performance),

CoNLL-U IO,

shared CoNLL-U library with MorphoDiTa and UDPipe,

perhaps structured NER.

NLP frameworks: Treex & Udapi

Martin Popel, Zdeněk Žabokrtský et al.

- **Treex** https://github.com/ufal/treex
  - Perl only, quite slow, no progress since ~2016
  - Treex::Web, support for tectogrammatical layer, TectoMT etc.

- **Udapi** http://udapi.github.io (API for UD, successor of Treex)
  - Perl, Java, **Python**
  - native support of Universal Dependencies and CoNLL-U
  - focus: speed and simplicity
  - format conversions and transformations
  - UD validity tests (ud.MarkBugs) and fixes (ud.ComplyWithText)
  - querying, data analysis (ud.See), evaluations (eval.Conll18)
  - taught in NPFL070 – Language Data Resources

Neural Monkey *an open-source toolkit for sequence learning*

Toolkit goals

- Code readability
- Modularity along research concepts
- Up-to-date building blocks
- Fast prototyping

Development

- Python 3.5, TensorFlow
- GPU support using CUDA and cuDNN
- Actively developed on Github

Achievements

- **317** stars on Github
- **20** Google scholar citations
- Significant amount of citations from *ACL conferences

`http://github.com/ufal/neuralmonkey`

Tensor2Tensor

Martin Popel et al.

- Transformer Neural MT model, T2T framework
- Best system in WMT2018 manual evaluation for English→Czech and Czech→English

  (significantly better than Google, Edinburgh etc.)
- See my PhD thesis for details

Sentiment analysis – what

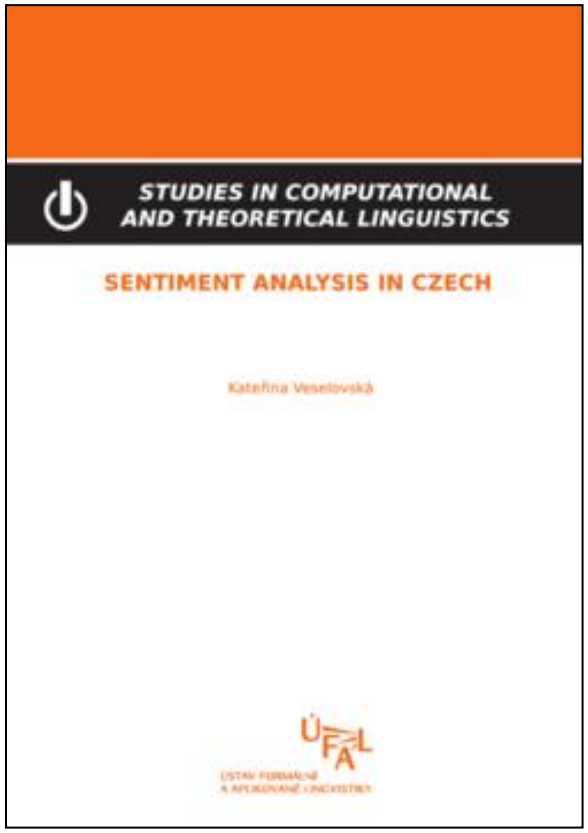
- automatic extraction of opinions or emotions given in a piece of text or in spoken data

  = what people actually think

Sentiment analysis – why

- product reviews
- public opinion surveys
- social media monitoring (personalized marketing)
- intention analysis (churn analysis)
- forensics (fraud detection)
- marketing trends prediction (stock/crypto market prediction)
- election outcome prediction
- healthcare applications

Sentiment analysis – results

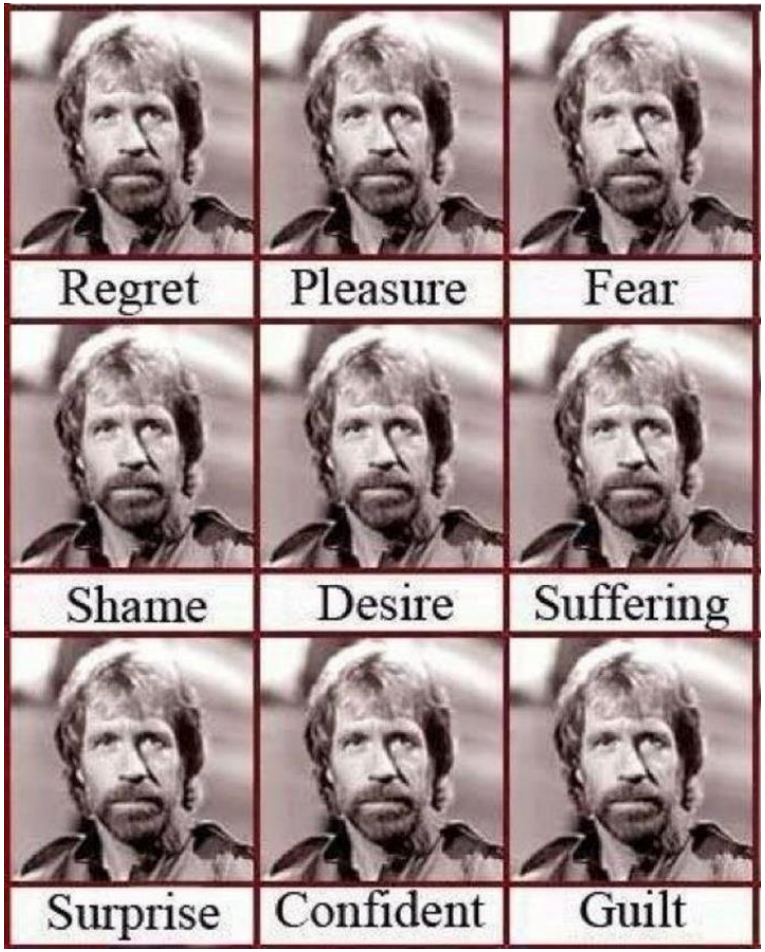


- **0,89** F-Score on positive/negative polarity detection
- **0,85** precision on opinion target identification
- one finished GAČR, several theses, endless hours of fun…

Sentiment analysis – next steps

- more experiments with neural networks & linguistics structure (diploma thesis V. Glončák)
- psycholinguistic experiments
- analysis of suprasegmental features
- multimodal data analysis

veselovska@ufal.mff.cuni.cz




LSD
in
3 minutes

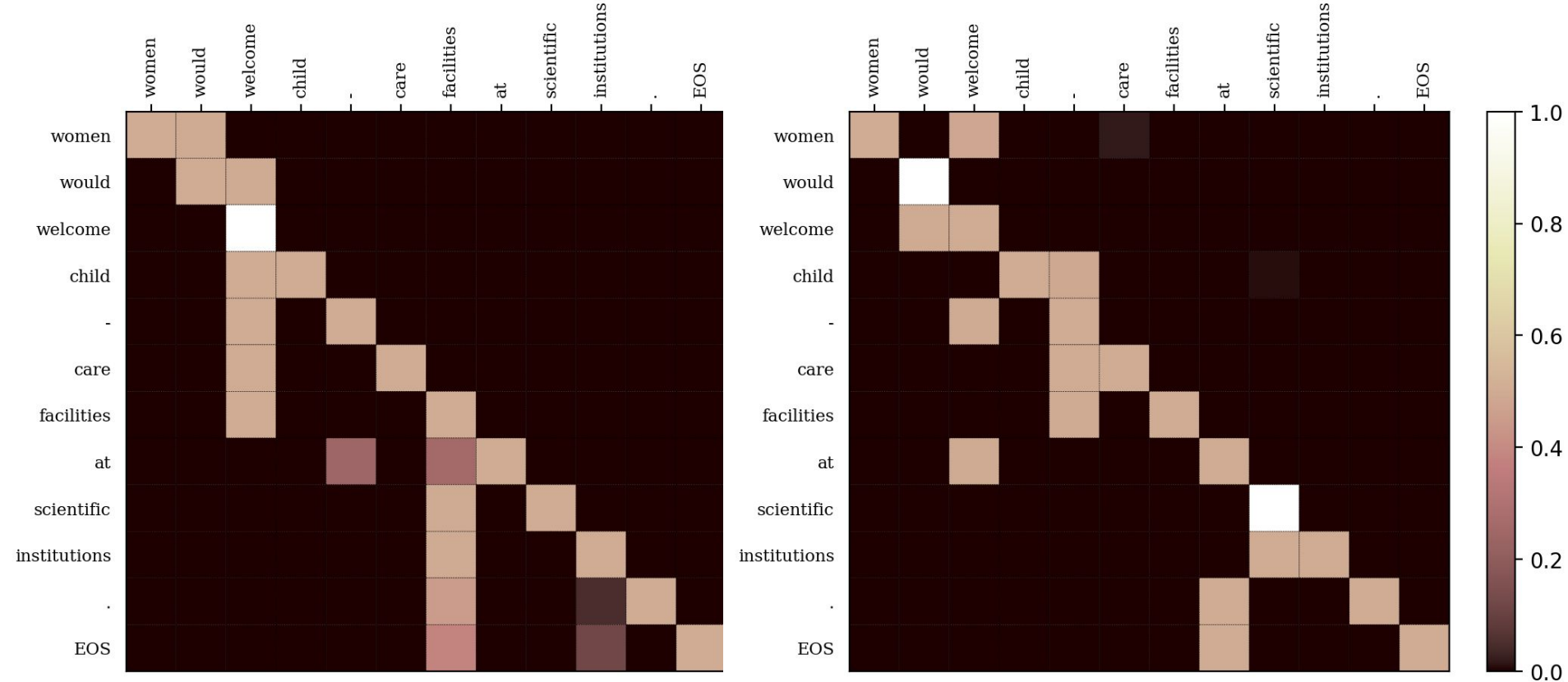
LSD - Linguistic Structure in Deep networks

- NLP tasks: machine translation, sentiment analysis, text summarization, ...
- solved by deep neural networks
  - in end-to-end fashion
  - with very little or no linguistic resources

Goals:

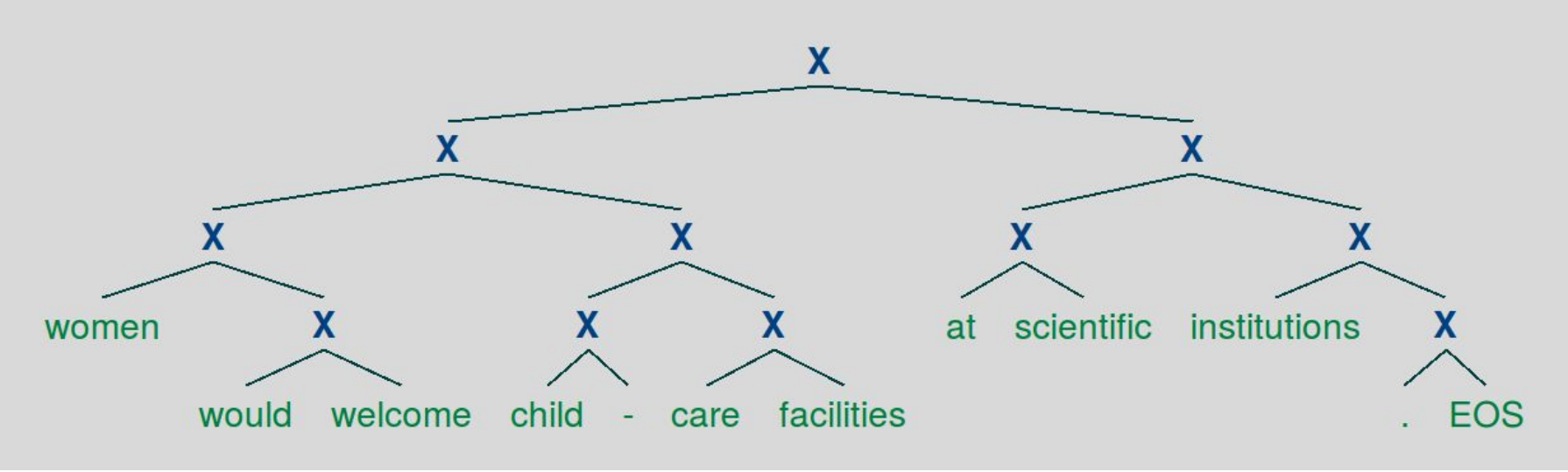
- To analyze NN, what specifically they learn inside for particular NLP tasks
  - What is learned in word embeddings, hidden states, attention mechanism?
  - Are there any linguistic features or structures we could compare to the established linguistic theories?
- We do NOT want to improve any of the NLP tasks

Example: Transformer encoder self-attentions



Example: Transformer encoder self-attentions (2)

… CKY applied across all its heads and layers:



Our team

- David Mareček
- Jindřich Libovický - Neural Monkey, ...
- Rudolf Rosa - morphology, syntax
- Tomáš Musil - semantic properties of word embeddings

Source code: ***http://github.com/ufal/lsd***

Online demo of NLP tasks: **http://quest.ms.mff.cuni.cz/neuralmonkey-czm/**

LINDAT/CLARIN becomes LINDAT/CLARIAH-CZ

Jan Hajič

Clarin ERIC

Since 2012 – Czech Republic founding member

19 members (countries) + 1 observer country

LINDAT CLARIN

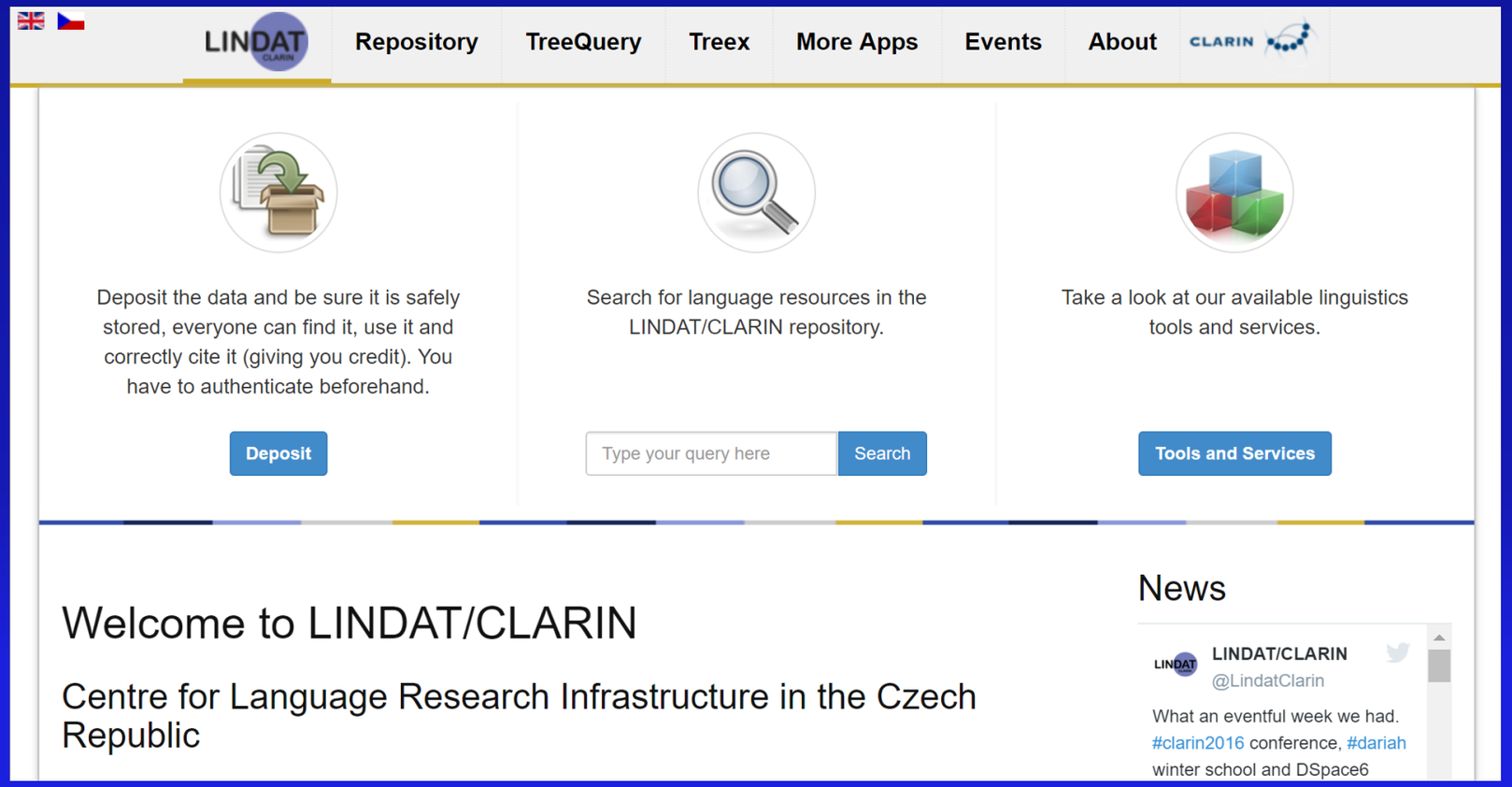
LINDAT/CLARIN

2016-2019

- Czech CLARIN “virtual” node, member of Clarin ERIC
- 1 center, 4 partners, 18+ people, 14M Kč/year budget
  - Charles University (Prague)
    - physical location of servers and data center
  - Institute of Czech Language, Academy of Sciences
  - Masaryk University of Brno
  - University of West Bohemia, Pilsen

LINDAT/CLARIN

- Data, tools and services at http://lindat.cz
  - Data: repository (dSpace, LINDAT developed), used in other countries
  - 19 major services & applications - all tools Open Source (incl. „IJP“)
    - Kontext (UD, parallel, speech), PML-TQ, lexicons
- Allows user-initiated deposit (i.e., for H2020 DMPs)
- 315 language resources or tools available (with Open Data) [253+62 ext.]



DARIAH ERIC

Contact

DARIAH-EU

ABOUT ACTIVITIES TOOLS & SERVICES NEWS – EVENTS

DARIAH in a Nutshell

Home > About

DARIAH ERIC: A network to enhance and support digitally enabled research and teaching across the Arts and Humanities

Description

The Digital Research Infrastructure for the Arts and Humanities (DARIAH) aims to enhance and support digitally-enabled research and teaching across the arts and humanities. DARIAH is a network of people, expertise, information, knowledge, content, methods, tools and technologies fr
maintains and operates an infrastructure in support of ICT-based res
using them to build, analyse and interpret digital resources. By worki
together individual state-of-the-art digital arts and humanities activit
preserves, provides access to and disseminates research that stems
best practices, methodological and technical standards are followed

DARIAH was established as a European Research Infrastructure Consortium (ERIC) in August 2014. Currently, DARIAH has 17 Members and several cooperating partners in eleven non-member countries.

Activity

DARIAH integrates digital arts and humanities research and activities from across Europe, enabling transnational

RECENT POSTS

DARIAH-FI Meeting in Helsinki
12 September, 2018

President of the Board of Directors
6 September, 2018

CAHIER Annual Workshop: Lexicography and Digital Editions

Digital Research Infrastructure for the Arts and Humanities

LINDAT CLARIN

LINDAT/CLARIAH-CZ

- Merge of CLARIN and DARIAH nodes in Czechia
- Timeline 2019-2022, budget ~170M CZK
  - 2019 in parallel with LINDAT/CLARIN
- Partners of Charles University:
  - Masarykova univerzita (FI, FF), Ústav pro jazyk český AV ČR, v.v.i., Západočeská univerzita v Plzni (FAV), Národní knihovna Praha, Moravská zemská knihovna, Národní galerie, Národní filmový archiv, Knihovna Akademie věd ČR, Filosofický ústav AV ČR, v.v.i., Historický ústav AV ČR, v.v.i.

LINDAT/CLARIAH-CZ Goals

- Provide digital data not only for language/linguistics, but for other [D]Hum: history, literature, philosophy, visual art, film, interdisciplinary fields w/humanities
- Services and tools (mostly language, but also multimedia)
- Provide access
- Multiple repositories (e.g. National Library + Library of the AV – KRAMERIUS system), LINDAT DSpace for rest
- Cooperate with both CLARIN ERIC and DARIAH ERIC
  - CZ will become full member of DARIAH ERIC
- Continue all CLARIN work: IJP, Center for Visual History Malach, etc.

LINDAT/CLARIAH-CZ Plans

- Expansion of data (PDT-C, lexicons, tools)
- Services (better MT, NE Linking, …)
- Integration in related projects (VIADAT)
- Any ideas?
  - LINDAT/CLARIAH-CZ exact plan yet to be written!

Model: czech-ud-1.2-160523

Input: (•) Plain text ( ) CoNLL-U ( ) Horizontal ( ) Vertical

Actions: [x] Tag and Lemmatize [x] Parse

A Input Text | Input File

I slovenský prezident Andrej Kiska vysekl tenistce Dominice Cibulkové poklonu po jejím triumfu na Turnaji mistryň. První slovenská vítězka vrcholu sezony i díky jeho reakci narychlo mění plány; ze Singapuru poletí ukázat slavnou trofej domů a teprve poté se vydá na dovolenou.

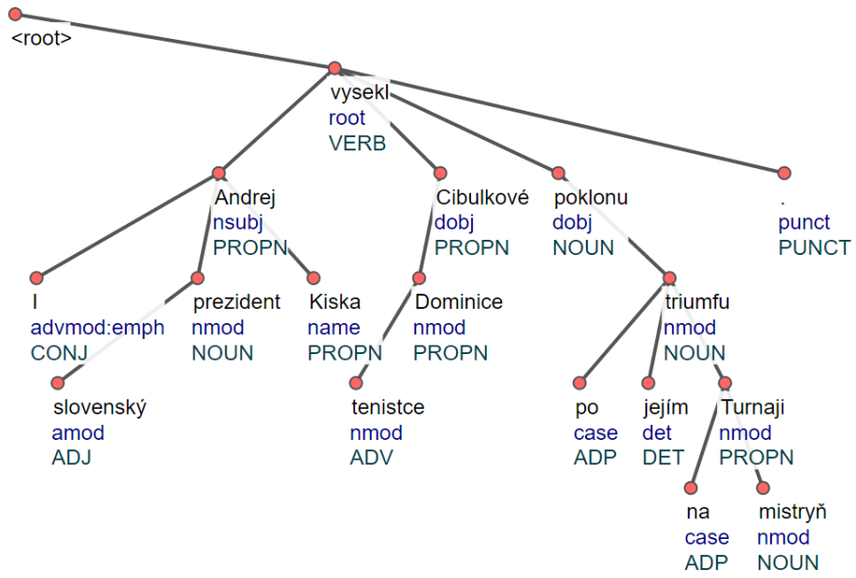
Process Input

A Output Text | Show Table | Show Trees

Save Tree as SVG

Previous 1 2 3 Next

I slovenský prezident Andrej Kiska vysekl tenistce Dominice Cibulkové poklonu po jejím triumfu na Turnaji mistryň .



Thank you!

http://lindat.cz

LINDAT CLARIN

CLARIN-LAPPS GRID interoperability Mellon Foundation Planning grant 2016-2018

Jan Hajič

(Pavel Straňák, Jozef Mišutka, Ondřej Košarko)

CLARIN – LAPPS Grid

Trans-Atlantic Interoperability
Connecting Language Processing Frameworks

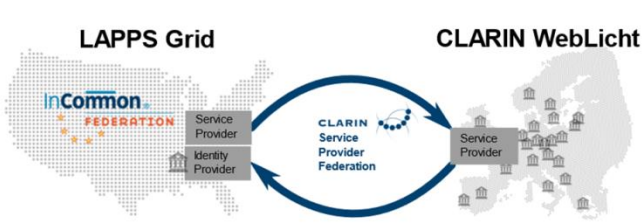
LAPPS Grid

- LAPPS Grid/Galaxy is used to create,

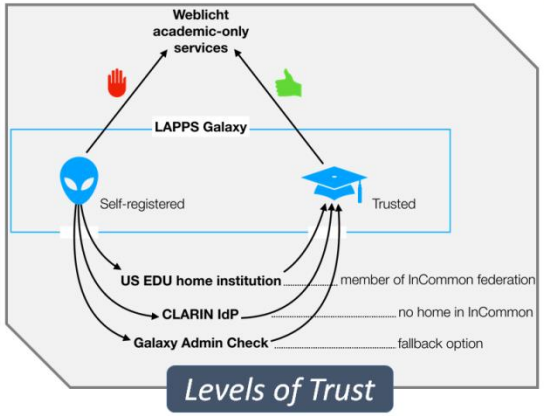
CLARIN/WebLicht

- WebLicht is used to create, execute,

AAI
Authentication Authorization Infrastructure



- CLARIN SPF includes more than 1700 EU institutions
- LAPPS users can increase Level of Trust via InCommon
- LAPPS Grid and WebLicht users can use services from both infrastructures



CLARIN-LAPPS Grid Continuation project?

- Implementation
  - AAI
  - Weblicht-Galaxy interoperability
  - Clarin Switchboard (resources to tools mapping)
  - LINDAT: AAI plus UDPipe,
    - new: Named Entity Linking (for news: en, de, cs)
- 2019-2021
- Still in preparation

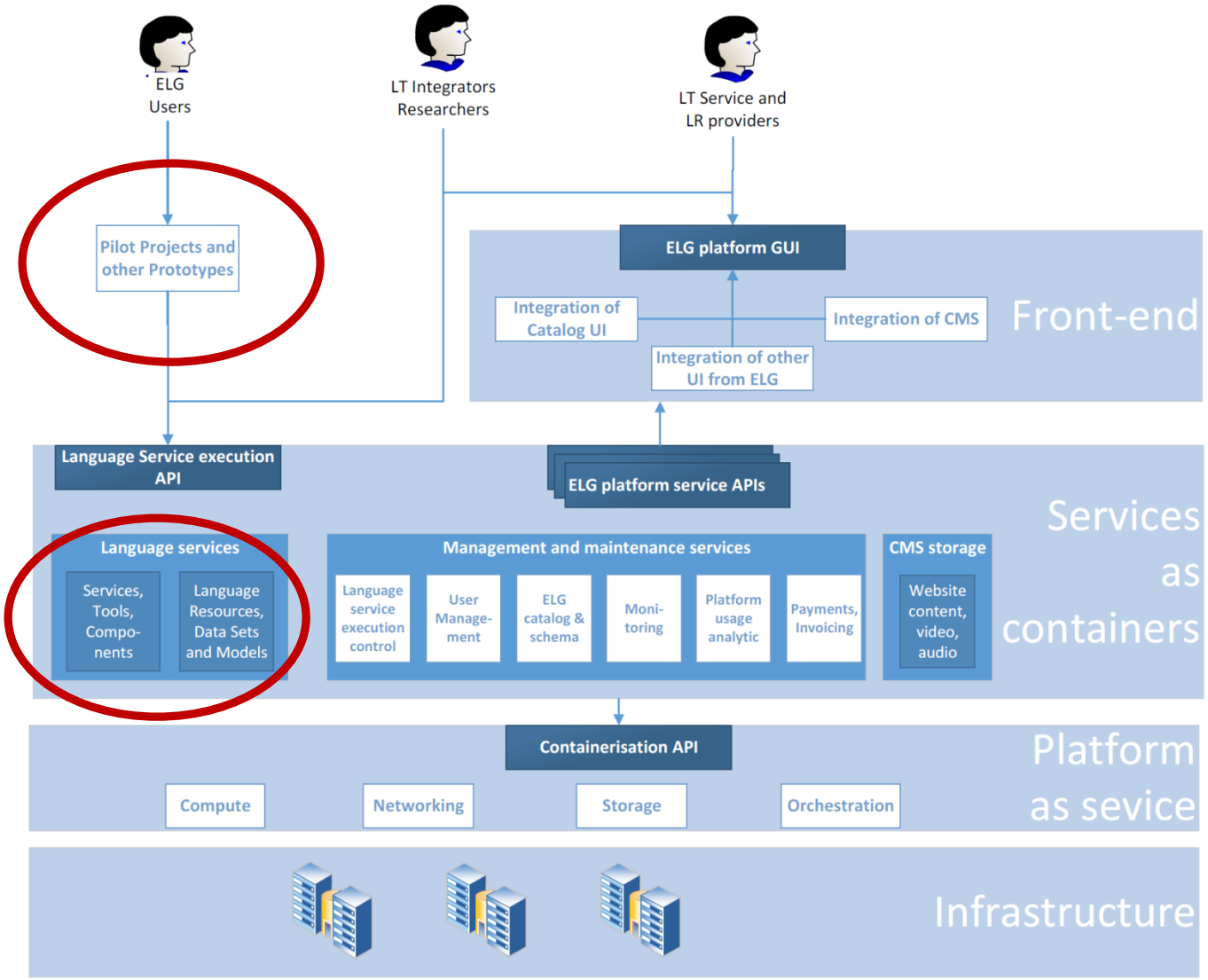
European Language Grid 2019-2021

Jan Hajič

ELG

- Platform for NLP resources, tools and services
- To be used in *commercial* applications
  - Research use for free
- Run by a company – business plan for future
  - Expected to run beyond project end
- DFKI, CUNI, Sheffield, Edinburgh, ELDA, Tilde, SAIL Labs, Expert System Iberia
- UDPipe might or might not be included

ELG structure



ELG budget

- CUNI:
  - 1,950,000 EUR to distribute to SMEs („3rd parties“)
  - 457k EUR for CUNI
  - 354K EUR labor
    - 84 PM (2,3 FTE), 24PM min. for Pilot management
    - ~12 PM overall mgmt
    - About ½ FTE for UDPipe and other tool integration
    - About ½ FTE for administration, organization of events

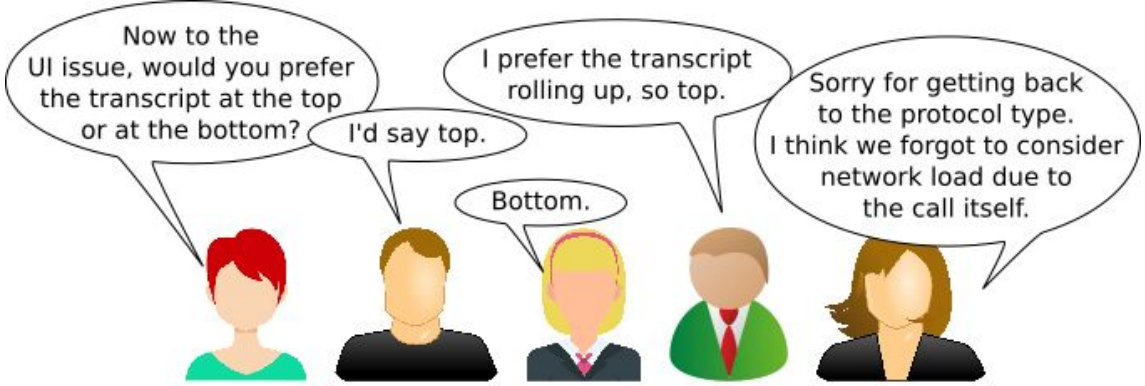
ELITR: European Live Translator

new! 2019-2021

Partners: CUNI (coord), UEDIN, KIT, alfaview, Pervoice, SAO (aka NKÚ)
Main goals:

- Machine interpreting at EUROSAI 2020 Congress in Prague (all EU+ SAIs).
- Document translation for audit reports of EUROSAI.
  - Highly multilingual: 24 EU languages + Albanian, Arabic, Armenian, Azerbaijani, Belorussian, Bosnian, Georgian, Hebrew, Icelandic, Kazakh, Luxembourgish, Macedonian, Moldovan, Montenegrin, Norwegian, Russian, Serbian, Turkish, and Ukrainian.
- (Speech translation in alfaview, something like GotoMeeting.)
- Automatic “minuting” of online meetings ≈ Speech&Dialogue Summarization:
  - Input: Bulleted agenda, speech during the call.
  - Output: Complete transcript, **Agenda populated with the minutes.**
  - Mode: off-line or optionally on-line (transcript and minutes grow in shared googledoc).
  - **We need you and your data!**
    - Whatever meetings you have from now on (Czech or English, or other langs.), please run them with our co-operation and donate the data.
    - (ELITR will handle GDPR+consent correctly, procedures in preparation.)

People: Ondřej Bojar, Anja Nedolužko, ...



Original agenda as prepared by the organizer beforehand:
- Protocol type: push or pull?
- Layout of the user interface:
  - Transcript grows at the top or bottom of the document?
  - Or in a side pane?

Shared document, everyone allowed to edit.
Starts with the agenda and gets populated by Automatic Minuting (AM)
- Protocol type: push or pull?
  - (AM) > Pull easier to implement.
  - (AM) > Updates can get lost with push *in case the user*
  - (AM) > Consider network load.
- Layout of the user interface:
  - Transcript grows at the top or bottom of the document?
    (AM) > Top (AM) > Bottom (AM) > Top, transcript rolling up.
  - Or in a side pane?

Transcript, optionally editable to correct ASR errors:
11:03 Sorry for getting back to the protocol type. I think we forgot ...
11:02 I prefer the transcript rolling up, so top.
11:02 Bottom
...

Bergamot: In-Browser Translation

new! 2019-2021

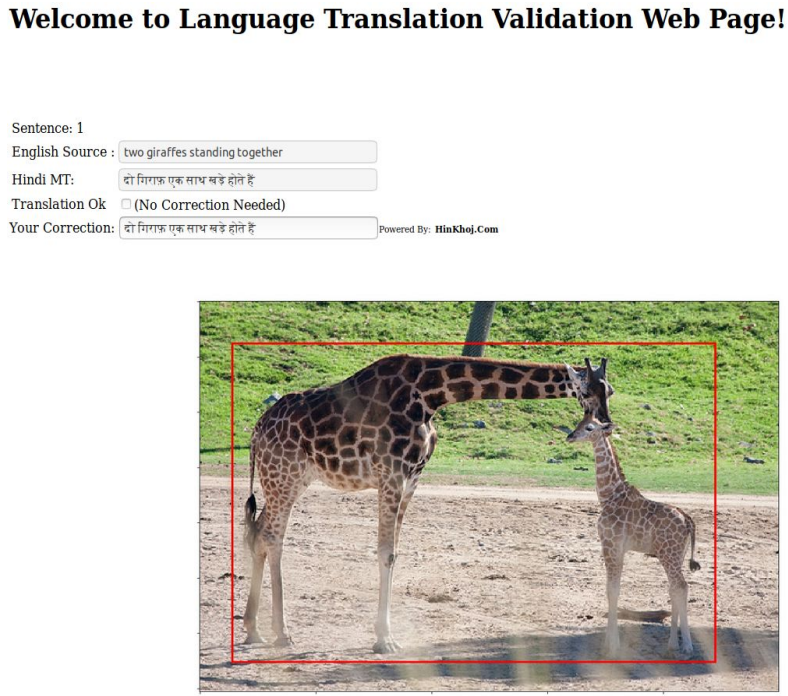
Partners: UEDIN (coord), CUNI, USHEF, UTARTU, Mozilla
Main goals:

- MT within browser
- Efficiency, efficiency, efficiency (model size, speed, non-GPU/low-end GPU).
- Quality estimation
- Dynamic adaptation
- **Outbound Translation** (“Gisting reversed”) <- our main
  - Design user interface and the underlying MT
    to allow reliably **produce** text in a foreign language
    given the unreliable MT engines.
  - Involves: quality/confidence estimation, source complexity estimation, automatic source simplification, many-source translation (“many-source” is having more variants of source, e.g. all the reformulations; technically related to multi-source).

People: Ondřej Bojar, Dušan Variš, Jindra Helcl, ...

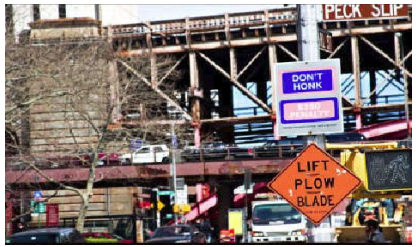
Hindi Visual Genome

- Development of a Dataset for English-Hindi Multi-Modal Machine Translation.
- 32K English-Hindi parallel segments, each being a label of a region in a picture.
- Test set: 1500 segments selected so that the image will be needed to disambiguate.
- English segments and images taken from “Visual Genome” (https://visualgenome.org/).
- Hindi Translation done by our NMT system (Tensor2Tensor) and post-edited by volunteers using our validation tool. http://ufallab.ms.mff.cuni.cz/~parida/index.html)



Searching for Ambiguous Words where Image Helps

- When translating the word “penalty”, the image can help:



But what other words are like that?

- Approach:
  - Cluster Hindi translations of English words based on their word embeddings.
  - Sort English words by level of separation of its Hindi clusters.
  - Manually validate.
- Result:
  - In the first 222 candidates, only 16 are sufficiently ambiguous.
  - In the next 222 candidates, only 3 are sufficiently ambiguous.
  - => Our ordering is reasonable, but such words are scarce.

Sl No.	Word
1	Penalty
2	Block
3	Press
4	Characters
5	Cross
6	Second
7	Fine
8	English
9	Players
10	Stand
11	Court
12	Stamp
13	Fast
14	Fair
15	Models
16	Date
17	Forms
18	Forces
19	Springs

Social Sciences & Humanities Open Cloud

SSHOC

Social Sciences and Humanities Open Cloud (SSHOC)

- where data, tools, and training are available and accessible for users of SSH data
- user-friendly tools & services
- links between people, data, services and training
- secure environments for sharing and using sensitive and confidential data
- multilingual aspects (sociological surveys …)

ÚFAL

#	Participant Legal Name	Country
1	CESSDA AS	NO
2	EUROPEAN SOCIAL SURVEY EUROPEAN RESEARCH INFRASTRUCTURE CONSORTIUM	UK
3	EUROPEAN RESEARCH INFRASTRUCTURE CONSORTIUM FOR THE SURVEY OF HEALTH, AGEING AND RETIREMENT IN EUROPE	DE
4	CLARIN ERIC	NL
5	DIGITAL RESEARCH INFRASTRUCTURE FOR THE ARTS AND HUMANITIES	FR
6	STICHTING LIBER	NL
7	KONINKLIJKE NEDERLANDSE AKADEMIE VAN WETENSCHAPPEN - KNAW	NL
8	UNIVERSITEIT VAN AMSTERDAM	NL
9	STICHTING KATHOLIEKE UNIVERSITEIT BRABANT	NL
10	TRUST-IT SERVICES LIMITED	UK
11	SEMANTIC WEB COMPANY GMBH	AT
12	FONDATION NATIONALE DES SCIENCES POLITIQUES	FR
13	THE UNIVERSITY OF NOTTINGHAM	UK
14	DEUTSCHES ARCHAOLOGISCHES INSTITUT	DE
15	CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE CNRS	FR
16	CONSIGLIO NAZIONALE DELLE RICERCHE	IT
17	UNIVERSITY COLLEGE LONDON	UK
18	FOUNDATION FOR RESEARCH AND TECHNOLOGY HELLAS	EL
19	STICHTING CENTERDATA	Netherlands

ÚFAL

CUNI (“linked third party”)

- WP 3: Lifting Technologies and Services into the SSH Cloud
  - T3.1 Multilingual Terminology
  - T3.3 Text & Data Mining
- WP 4: Innovations in Data Production
  - domain specific (social surveys) MT system for CAT
- 28,5 PMs, 40 months, from January 1, 2019

ÚFAL