Self-presenting slides

Charles University in Prague

Institute of Formal and Applied Linguistics

Prčice, September 14 & 15, 2017

Petra Barančíková

just one more,
I promise! :)

Dissertation topic: **Paraphrasing for Machine Translation Evaluation** (5th year)

Recent past:
- Internship at Google
- ParaDi with Vendula Kettnerová

Present/Future:
- Work at Seznam.cz
- Dissertation
- 3rd SloNLP with Ruda Rosa
- fun side project: Receptron

Eduard Bejček

- **PDT-C** (consistency checks of **morphology** annotation and lexicon) → data management, annotators
- **Subfunctors** (categorization of **LOC** & **DIR***) → thousands of examples extracted from PDT-C and presented in elaborate interactive table
- LaTeX, **SCTL** (ÚFAL publishing house) → compatible templates for both **PhD thesis** & SCTL "orange book"
- **Named Entities** (in **PDT-C**) → yesterday's presentation
- **Multiword Expressions** (international effort in 18 languages: methodology, corpus, **PARSEME** shared task)
- **Valency** (VALLEX web pages) → ufal.cz/vallex/3.0/guide.html

- **ÚFAL Beer Committee Founding Member** → beer (Oct 12)

Ondřej Bojar

Topics: everything around MT

- Utilizing linguistic analyses in MT.
- Document-level translation.
- Interpreting what NMT is learning.

Events: MT Marathons, EAMT2017, WMT

- MT Marathon 2018 again foreseen at UFAL.
- WMT18 to include doc-level eval and error analysis.

Projects: research (HimL, QT21), coordination (CRACKER)

- New EU call coming out soon ($\sim$Oct$\rightarrow$Mar).
- Searching for technical writers.

Anyone can help spending $675117 Azure credits by Oct 10?

Libuše Brdičková

Sekretářka ÚFAL – IV. p. , č. dv. 408

Pracovní doba: 7:30 – 16,00

Středa: 7:30 – 8:30 děkanát, dále na MS

Malostranské nám. 25

118 00 Praha 1

Libuše Brdičková

1. Evidence, zpracování CP (zálohy na cestu, vyúčtování)

2. Návrhy na přijetí zahraničních hostů (zálohy, vyúčtování)

3. Sledování rozpočtu 207-01/PROVOZ, SVV, PROGRES, studentské projekty (A. Abrehimian, T. Kocmi, M. Vodolán, K. Droganova, N. Mediankin), běžná hospodářská agenda, zpracování faktur, plateb do zahraničí, vyřizování objednávek všeho druhu

4. Příprava obhajob DP, SZZ

5. Evidence docházky

Libuše Brdičková

6. Osobní kontakt s děkanátem (hospodářské odd., stud. odd.)

7. Vyúčtování záloh (stálé, mimořádné)

8. Realizace plateb platební kartou

9. Zásobování pracoviště základními kancelářskými potřebami, kávou atd., vybavení lékárničky

10. Evidence a objednávání stravenek

Karry

Karolína Burešová

- To-be 1st year Ph.D. student
- Supervised by Pavel Pecina

- Main topic: Text simplification
- Related: Multi-word expressions, coreference, paraphrasing, language modelling
- Making use of: Morphological analysis and generation, parsing

Text simplification: basic idea

This thesis researches text simplification, focusing on Czech, a Slavic language, offering various approaches to some simplification subproblems (albeit the simplification problem is solved neither thoroughly nor as a whole), thus shedding some light on a problem of non-negligible importance for several target groups of notable sizes.

$\rightarrow$

This thesis deals with text simplification. It works with Czech (a Slavic language). It doesn't solve simplification completely but it tries to solve some of simplification tasks. Text simplification can be important for many different people.

My current work aimed at "simple (imperfect) Czech" native speakers

Silvie Cinková

Reviving Zellig S. Harris: more syntactic information for distributional semantics (GAČR grant 2015-2017)

- What makes two ~~lexicon senses~~ usage patterns prone to interannotator confusion in WSD? (Corpus Pattern Analysis)
  - correlation of graded annotator decisions with syntax, entailment, distr. similarity of arguments, factuality

  with Anna Vernerová and Ema Krejčová
- Do various linguistic transformations improve the performance of a distributional semantic model/embedding model? English, Czech
  - in particular morphological derivations

  with Vincent Kríž and Iveta Kršková

Silvie Cinková

- Linguistics with data analysis in R
  - learning
  - teaching
  - helping
  - simple statistical methods, advanced data-wrangling and graphing libraries, string processing

(ggplot2, dplyr, tidyr, stringr)

with Václav Cvrček and David Lukeš
from the Institute of the Czech National Corpus, Faculty of Arts

Silvie Cinková

- Language Intelligibility Awareness
  - UN Convention on the Rights of Persons with Disabilities (2006) includes plain language
  - Legislative Drafting Guide (2015, http://eur-lex.europa.eu/content/techleg/EN-legislative-drafting-guide.pdf),
  - comparison of the syntactic differences between written standard vs. administrative language across languages vs. ***plain language*** (English, Scandinavian languages..., Japanese)
  - plain language easier for MT? like "controlled language"?

Silvie Cinková

- help with project proposals & reports
  - preliminary research for "State of the Art" sections
  - proofreading
  - translations
- member of the executive board of the Czech Association for Digital Humanities
- member of the editorial board of *Orð og tunga*

Mgr. Ing. Kira Droganova (droganova@ufal)

- Universal morphosyntactic annotation of language data (Univerzální morfosyntaktická anotace jazykových dat).

- UD Russian SynTagRus.

Universal morphosyntactic annotation of language data

... It's about non-trivial syntactic trees

Project tasks:

- to examine existing theories and annotation standards
- to collect and prepare the data where elliptical constructions can be extracted from
- to propose modified or improved method of annotation
- to explore parsing and learning tools and algorithms applied to the prepared data
- to develop a novel method?

UD Russian SynTagRus

SynTagRus treebank of Russian

- Meaning—Text Theory
- 1 MW
- high granularity (67 syntactic relations)
- Corpus search: http://ruscorpora.ru/en/search-syntax.html

- Data quality
- UD Russian SynTagRus & UD Russian

Thank You!

Petra Galuščáková

- **Information** and **multimedia** retrieval

- Multimedia
  - Retrieval and linking video segments
  - Query text and query segment
  - Combination of lexical, visual and audio features
  - SHAMUS (UFAL Search and Hyperlinking Multimedia System)

Multimedia Retrieval

- 4000 hours of BBC video broadcast (MediaEval and TRECVid Benchmarks)
- Subtitles and automatic transcripts
- Visual information
  - Feature Signatures (KSI, Siret Group)
  - Caffe descriptors (DISA, MUNI)
  - Face descriptors (CMP, CTU)
- Audio information
  - Prosodic features
  - Music

Information Retrieval

- Text entities and relations retrieval
- Malach
  - Czech Malach Cross-lingual Speech Retrieval Test Collection
- Digital Editing of Medieval Manuscripts

Jindřich Helcl

Main topic: **Neural Machine Translation**

Research

- Multimodal Translation (joint work with J. Libovický)
  - Attention Strategies for Multi-Source Sequence-to-Sequence Learning (ACL '17)
  - Submissions to WMT shared tasks
  - More fine work coming up!
- New English-Czech dataset for MMMT task next year
- Co-organizing WMT17 Neural Training Task (with OB, JL, TK & TM)
- **Neural Monkey** toolkit (with JL, TK, DV and others)
  - Use the monkey! `github.com/ufal/neuralmonkey`

Teaching

- NPFL116 – Compendium of Neural Machine Translation
  - Together with J. Libovický
  - Too much free time? Sing up for our course! `ufal/courses/npfl116`

Jarka Hlaváčová

- Czech morphology- updates of dictionary:
  - revisions, error fixing,
  - new words,
  - checks,
  - morphological service (e.g. derivational relations)
- Cooperation with ÚTKL, ÚČNK, new morphology
  - categories revisited, some new values
    - e.g. new POS „foreign word“ - already implemented in MorfFlex

Personal profile

Vojtěch Hudeček

September 13, 2017

Sedlec-Prčice

Education

- Faculty of Mathematics and Physics
  - Bachelor's degree in General Computer Science
  - Master's degree in Artificial Intelligence

Student works

- **Bachelor thesis** – Distributed video compression using peer-to-peer network
- **Master thesis** – Improving pronunciation of TTS systems, based on user's recordings

Interests

- Automatic Speech Recognition and Speech synthesis
- Dialogue management
- Artificial Neural networks

Me at ÚFAL

- supervisor Zdeněk Žabokrtský
- extension and modificiation of **the Derinet**
- exploring unusual neural networks architectures and its applications in NLP

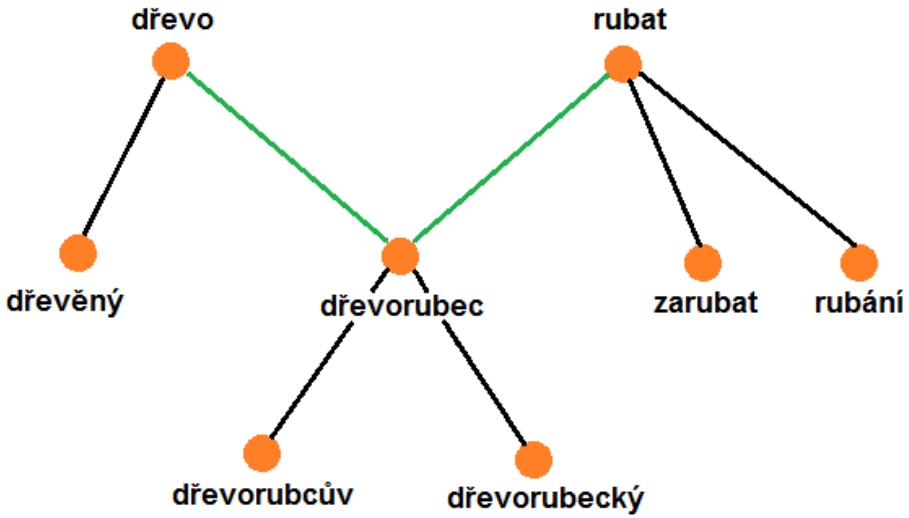
Thank you

DELIMITATION OF COMPOUNDS

Adéla Kalužová

BASIC INFORMATION

- 1st year Ph.D.
- supervisor: Mgr. Magda Ševčíková, Ph.D.
- topic: Formal Representation of Compounding
- background: DeriNet database



CURRENT SITUATION

- about 30 000 potential compounds identified and checked manually
- different groups – which should we consider actual compounds?

EXAMPLES OF GROUPS

- clear cases: *velkovýroba* (large + production)
- one part present, the other missing in DeriNet (not a full-meaning PoS):
  čtyřdveřový (four + door + adj. ending);
  DeriNet only contains N, V, Adj, Adv
- neoclassical: *kardiologie* (both parts in DeriNet) but *psychologie* – only second part
- originally compound loan words: *biftek*, *gólman*
- abbreviations: *Čedok*, borderline: *pančelka*
- “false” compounding: *monokiny* (an. *bikiny*)
- duplicate: *jistojistý* (sure + sure = very sure)

FUTURE PLANS

- further compound identification
- parent identification (splitting)
- formal representation (modification of DeriNet structure)

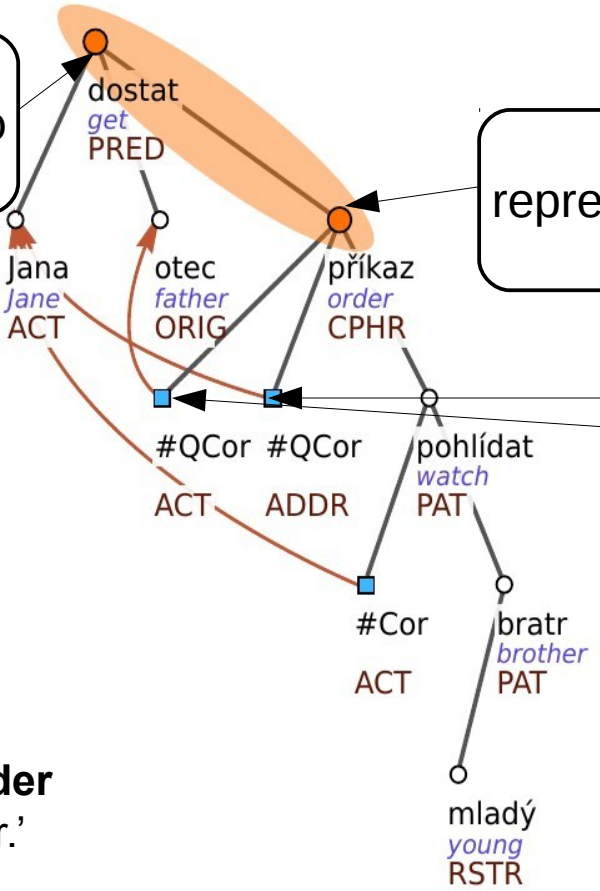
Václava Kettnerová

2015-present Combining Words: Syntactic Properties of Czech Multiword Expressions with Light Verbs, supported by the GAČR, with Markéta Lopatková, Petra Barančíková & Eda Bejček
LINDAT-Clarin

Representation of Czech light verbs

PRED
representing the light verb
ACT CPHR ?ORIG

CPHR
representing the predicative noun
ACT ADDR PAT



coreference

Jana dostala od otce příkaz pohlídat mladšího bratra.

'Jane **got** from father the **order** to watch her younger brother.'

syntactic structure

~ **impf: poskytovat pf: poskytnout** [blu-v-poskytnout-poskytovat-2]
+ ACT(1;obl) ADDR(3;obl) CPHR(4;obl)
-note: (diat) **impf:** *deagent* %zákazníkům se poskytují kvalitní služby% [made-up]
passive-být %Závody navštěvuje mnohem víc diváků, tedy i naše služby jsou poskytovány ve větším rozsahu.% [automatic]
passive-bývat %Roční servis výrobci většinou poskytují zdarma, přičemž zprostředkovatelem bývá poskytován za úhradu.% [automatic]
pf: *deagent* %zákazníkům se poskytly kvalitní služby% [made-up]
passive-být %Při prohlídce v hodnotě 298 Kč budou navíc překontrolována světla a bude poskytnuta tříměsíční asistenční služba zdarma.% [automatic]
passive-bývat %Pokud však dárce finanční odměnu vyžaduje, bývá mu poskytnuta.% [automatic]
poss-result-nconv-mít %Podle aktuálních čísel má Komerční banka poskytnuty úvěry za 244 miliard korun.% [SYN]
poss-result-conv-mít
-full: blu-v-poskytnout-poskytovat-1
-lvc1: dotace, informace, interview, pomoc, půjčka, rozhovor, záruka
-map1: ACTv-ACTn, ADDRv-ADDRn
-example1: **impf:** Řada ministerstev poskytovala dotace neziskovým organizacím v rámci svých vyhlašovaných programů.;
Jako cizinec žijící v Zimbabwe mi poskytuje informace, které bych asi od místních lidí jen stěží získal.;
Neposkytovala jim ani tak interview, jako spíše předváděla proces dýchání.;
Bůh poskytuje pomoc Svou tomu, komu chce.;
Většina bank soudila, že poskytovat malé půjčky průměrným rodinám nestojí za čas ani úsilí.;
Mezi jiným i tím, že vám teď poskytuji rozhovor.;
Podle Bohuslava Sobotky by nebylo zodpovědné poskytovat další záruku za tak velký úvěr.
pf: Ministerstvo financí již kraji poskytlo dotaci patnáct milionů korun na pokračování prací ve skladu.;
Předpokládám, že vaše oddělení poskytne informace o tom zatčení tisku.;
Jako manželka velícího generála poskytla interview listu Washington Post den po úspěšném vylodění v Maroku v roce 1942.;
Jakou ti, prosím tě, může poskytnout pomoc?;
Dodal, že město poskytne půjčky všem deseti zájemcům v průměrné výši v rozmezí 250 až 300 tisíc korun.;
Pouhý měsíc před smrtí poskytl rozsáhlý rozhovor reportéru Mladého světa Romanu Lipčíkovi.;
Střešní krytinu i ostatní materiál budeme vždy raději kupovat u prověřených a certifikovaných prodejců, kteří nám na něj poskytnou zákonnou záruku.
-lvc2: ochrana, opora, péče, podpora, služba, ubytování, vzdělání
-map2: ACTv-ACTn, ADDRv-PATn
-example2: **impf:** Rozárce větve košatého dubu poskytovaly dokonalou ochranu, a proto ji tam nikdy nikdo neviděl.;
Stanovení cíle je systémovým prvkem v pojetí předmětu a jeho jasné formulování spolu s kontinuitou s konkrétním obsahem předmětu poskytují učiteli oporu.;
Zařízení bude poskytovat komplexní celoroční péči zhruba 120 lidem.;
Obě náboženství jim poskytují duchovní podporu a umožňují žít v rovnováze.;
Zakládám firmu, která by venčila psy a poskytovala další služby.;
Nyní poskytuje ubytování hráčům Karlovarského symfonického orchestru.;
Německé školy poskytovaly vzdělání v německém jazyce.
pf: Každému člověku je nutné poskytnout ochranu, když někdy upozorní na nekalou činnost.;
Mohla poskytnout oporu sestře i mamince.;
Tvrdí, že jejich matce tamní zdravotníci neposkytli patřičnou péči.;
Generální ředitel skupiny VW Martin Winterkorn ujišťuje, že úřadům při kontrole aut poskytne absolutní podporu.;
Organizátoři chtějí poskytnout služby minimálně třem tisícovkám nezaměstnaným.;
Rodiny v Kolíně a okolí poskytli ubytování devadesáti tisícům poutníků.;
Navštěvoval školy a také synovi poskytl vzdělání.
-lvc3: obživa, potěšení, útěcha
-map3: ACTv-PATn, ADDRv-ACTn
-example3: **impf:** Hlohyně šarlatová představuje dokonalý úkryt pro ptačí hnízda, opeřencům poskytuje i hojnou obživu v zimě.;
A nakonec z ní zase zbývá pouze báseň, která stále ještě mnohým poskytuje potěšení, i když se z nějakého důvodu nedostává do antologií.;
Jeho dopisy mi poskytovaly útěchu.
pf: Jedinou obživu ve vězení mu poskytly krysy. [made-up];
A pak mi ta práce poskytla velké potěšení.;
Samotné pokleknutí o samotě mezi studenými přísnými kameny a skoro němé pronášení důvěrně známých slov mu poskytlo větší útěchu a ujištění, než se odvažoval očekávat.
-lvc4: možnost, obživa, potěšení, příležitost, útěcha
-map4: ADDRv-ACTn
-example4: **impf:** Členitý terén na ostrově poskytoval rozsáhlé možnosti obrany.;
Řeka jim poskytuje obživu a je pro ně druhým domovem.;
Neposkytujme našim nepřátelům potěšení radovat se z hrůzy, kterou po světě šíří.;
Zmíněný příklad mi však poskytuje příležitost zamyslet se nad otázkou zásadnější.;
Rodina je mu, pokud možno, oporou a poskytuje mu útěchu.
pf: Prudké výkyvy poskytly vědcům možnost hodně přesného měření.;
Chudá půda byla sotva schopná poskytnout zemědělcům obživu.;
Pstruzi se zachovali sobecky, vůbec nám neposkytli potěšení vytáhnout je z vody.;
Ale otec mu neposkytl příležitost ke smíření.;
Musí být schopni poskytnout duchovní útěchu stejně tak křesťanům jako například muslimům, židům či hinduistům.
-instig4: ACT

- 1025 complex predicates with light verbs
- 129 verb lemmas of light verbs
- 560 nouns
- 16 types of coreference

VALLEX

Paraphrasing of complex predicates with light verbs by single verbs

with Petra Barančíková

Tom Kocmi (kocmi@ufal) starting 3rd year PhD

- Topic: Neural Machine Translation
  - Thesis: Document Embeddings as a Mean of Domain Adaptation
  - Supervisor: Ondřej Bojar
- Side research:
  - Language Identification (EACL 2017)
  - Word Embeddings (word2vec)
  - Document Level MT
  - Multi-task learning
  - Summarization
- Developing: Neural Monkey
- Co-organizing: WMT17 Training Task, EAMT 2017

Matyáš Kopp

- **PML Tree Query** and related tools
  - PMLTQ Perl core module
  - PML-TQ Sever
  - PML Tree Query Interface for TrEd
  - PML-TQ Web interface
- euler.ms.mff.cuni.cz administration and data management
- PML-TQ technical user support

kopp@ufal

Matyáš Kopp

- Colaborants: Pavel Straňák, Jiří Mírovský, Daniel Zeman, Anna Vernerová
- Supported by LINDAT/CLARIN project of the Ministry of Education of the Czech Republic (project LM2015071)

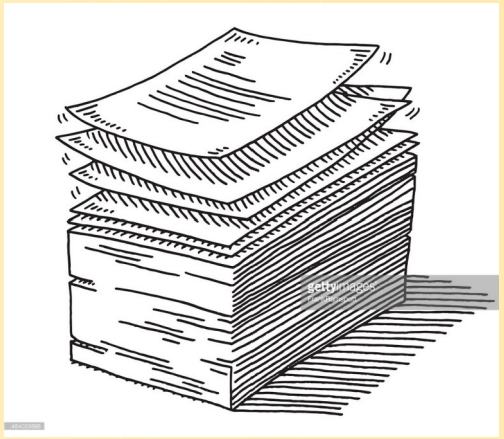
kopp@ufal

Administration staff

Project managers

Marie Křížková, Kateřina Bryanová, Jana Hamrlová

Marie Křížková (since 1999)



- Maintaining records of job positions on all projects in ÚFAL
- Maintaining and monthly check-up of all wages paid in ÚFAL (calculation of personnel costs balance, consultation of personnel costs with investigators of all Czech projects, preparing bonuses and job contracts for Czech projects)
- Czech projects: all projects (except of Viadat) of prof. Hajič (e.g. LINDAT, NAKI ÚSTR), GAČR (CEMI) of P. Pecina, support for other investigators
- Administrating of Industry Cooperation (invoicing, financial drawing)

Kateřina Bryanová (since 2011)

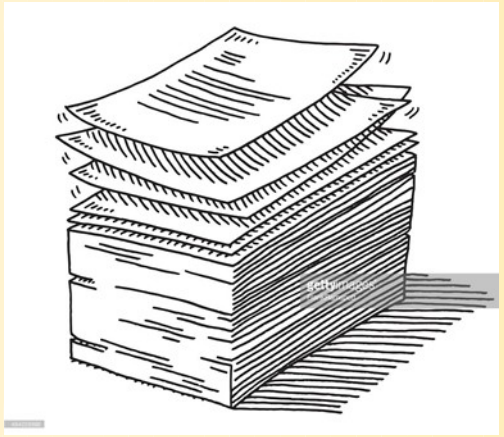
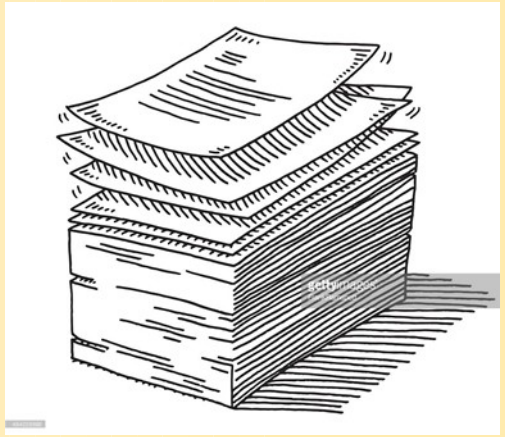
Project manager: administration, communication with the financial providers, financial drawing, invoicing, maintaining costs balance, personnel costs, administrating bonuses and job contracts,…

EU projects: HimL, CRACKER, QT21, CLARIN plus

DigiLing, Mellon Grant, Clarin Secondment

Czech projects: NAKI VIADAT

Jana Hamrlová
(since July 2017)

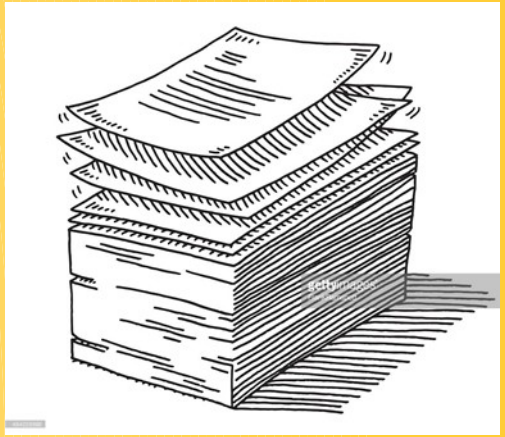




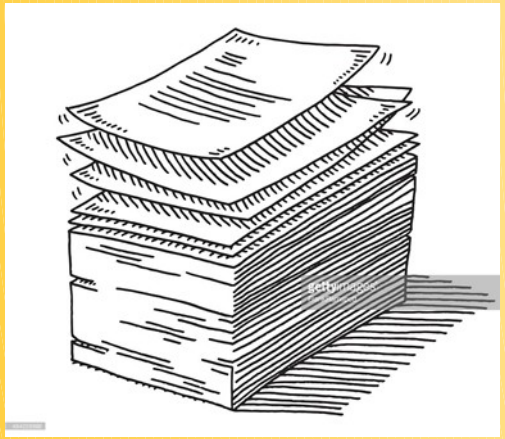
Project manager: administration, communication with the financial providers, financial drawing, invoicing, maintaining costs balance and personnel costs, administrating bonuses and job contracts,…

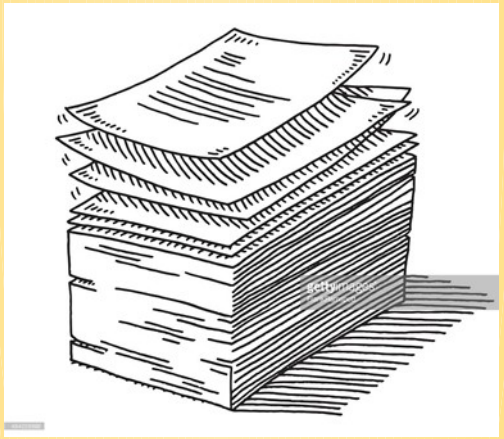
OP VVV projects: LINDAT, LangTech

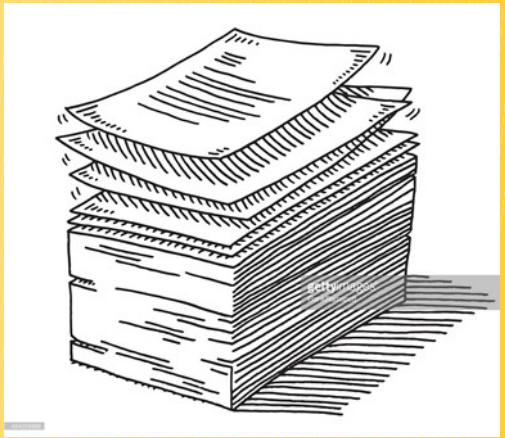
OP PPR projects: OP PPR 1 translation, OP PPR 3 document

Thank you for your attention









Oldřich Krůza: Radio Makoň

- Topic: Iterative transcription system exploiting listeners' feedback

- Ph.D. study commenced: Oct. 2011
- Interrupted: Oct. 2014 – Sept. 2017

Oldřich Krůza: Radio Makoň

Material

Volume: 1000+ hrs. of recordings
Single speaker: Karel Makoň
Single topic: mystic
Varying quality

Oldřich Krůza: Radio Makoň

Previous Work

- Acquisition of automatic transcription
- Prototype of a web application for correcting the transcription

Oldřich Krůza: Radio Makoň

Work during the time off

- Maintenance and minute enhancements
- Search
- Normalizing MFCCs on isolated utterances
- Rewrite of the web application

Oldřich Krůza: Radio Makoň

Work during the time off:

Search

- Elastic
- Stemming Czech (rule-based wins)
- Searching by phonemes

Oldřich Krůza: Radio Makoň

Work during the time off:

Normalizing MFCCs

- Attempt better normalization than HTK does out of the box
- Cutting off utterances only (filtering out `sp`, `sil`)
- Low-level processing MFCCs with Perl

Oldřich Krůza: Radio Makoň

Web App Rewrite

- Technology update necessary
  - Flash is dead
- Targeting both the community and public
- Optimize for sharing on social networks
- Technology used:
  - Web standards
  - React / Redux
  - Bootstrap

Oldřich Krůza: Radio Makoň

Look-ahead

- Finish new web front-end
- Employ neural networks in acoustic model
- Engage public
  - Topic indentification
  - Better search
  - Organic recruitment of transcribers

Markéta Lopatková – Research Projects

Research interests / research projects:

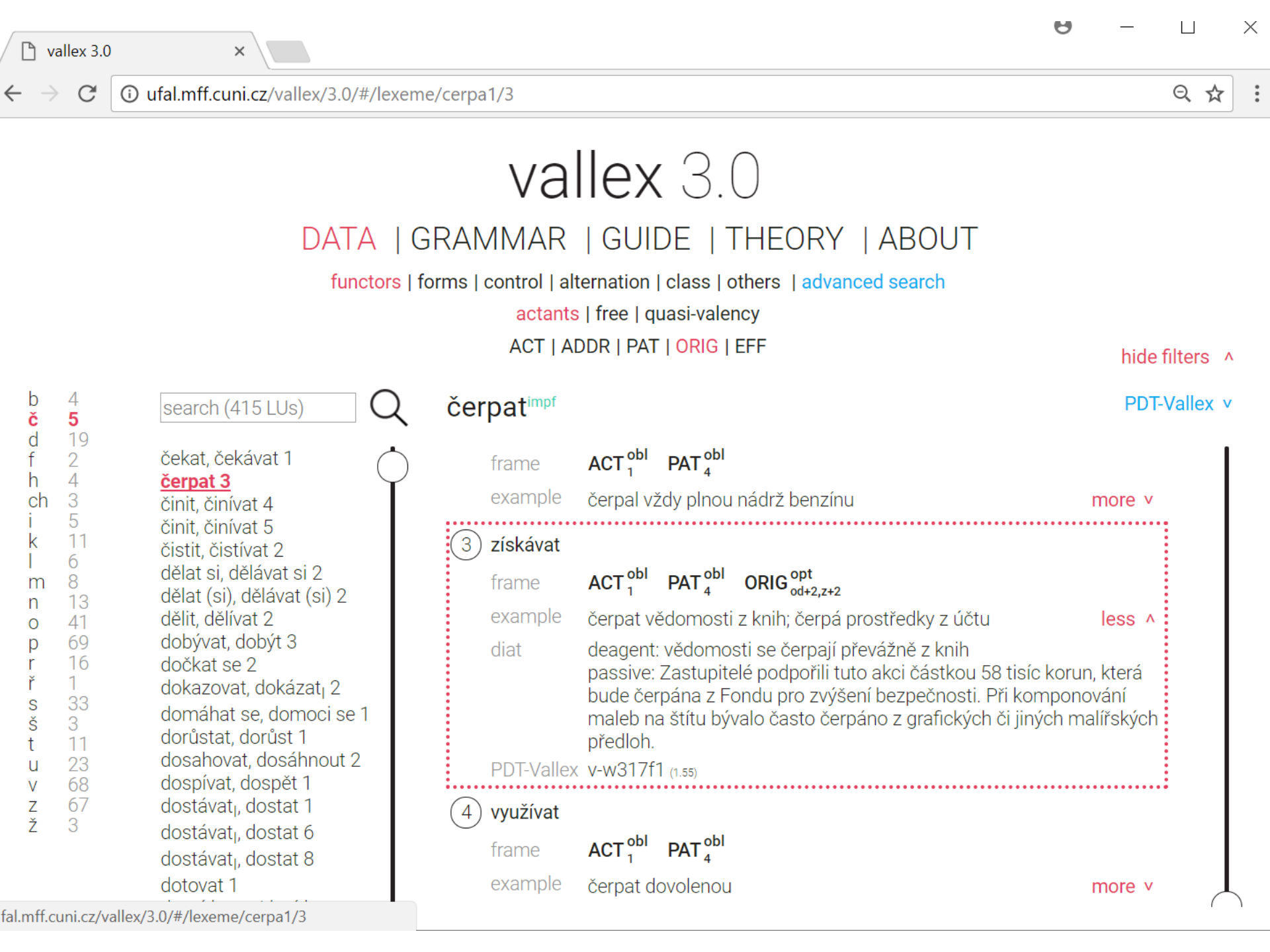
- Valency lexicon of Czech verbs – VALLEX
  with Václava Kettnerová, Anša Vernerová, Eda Bejček, Petra Barančíková
  (past - Zdeněk Žabokrtský)

- Modeling of stratificational dependency-based syntax
  based on the analysis by reduction and restarting automata
  esp. with Martin Plátek (KTIML – Department of Theoretical Computer Science and Mathematical Logic)

Markéta Lopatková – Research Projects

Valency lexicon of Czech verbs – VALLEX

- changes in valency structure of verbs, their representation in a lexicon
- Delving Deeper: Lexicographic Description of Syntactic and Semantic Properties of Czech Verbs, GAČR 2012-15(-17)
- http://ufal.mff.cuni.cz/vallex/3.0/


vallex 3.0
ufal.mff.cuni.cz/vallex/3.0/#/lexeme/cerpa1/3
vallex 3.0
DATA | GRAMMAR | GUIDE | THEORY | ABOUT
functors | forms | control | alternation | class | others | advanced search
actants | free | quasi-valency
ACT | ADDR | PAT | ORIG | EFF
hide filters
b 4
č 5
d 19
f 2
h 4
ch 3
i 5
k 11
l 6
m 8
n 13
o 41
p 69
r 16
ř 1
s 33
š 3
t 11
u 23
v 68
z 67
ž 3
search (415 LUs)
čekat, čekávat 1
čerpat 3
činit, činívat 4
činit, činívat 5
čistit, čistívat 2
dělat si, dělávat si 2
dělat (si), dělávat (si) 2
dělit, dělívat 2
dobývat, dobýt 3
dočkat se 2
dokazovat, dokázat| 2
domáhat se, domoci se 1
dorůstat, dorůst 1
dosahovat, dosáhnout 2
dospívat, dospět 1
dostávat|, dostat 1
dostávat|, dostat 6
dostávat|, dostat 8
dotovat 1
čerpat impf
PDT-Vallex
frame ACT 1 obl PAT 4 obl
example čerpal vždy plnou nádrž benzínu
more
3 získávat
frame ACT 1 obl PAT 4 obl ORIG od+2,z+2 opt
example čerpat vědomosti z knih; čerpá prostředky z účtu
less
diat deagent: vědomosti se čerpají převážně z knih
passive: Zastupitelé podpořili tuto akci částkou 58 tisíc korun, která bude čerpána z Fondu pro zvýšení bezpečnosti. Při komponování maleb na štítu bývalo často čerpáno z grafických či jiných malířských předloh.
PDT-Vallex v-w317f1 (1.55)
4 využívat
frame ACT 1 obl PAT 4 obl
example čerpat dovolenou
more
fal.mff.cuni.cz/vallex/3.0/#/lexeme/cerpa1/3

vallex 3.0

1.5 Recipientní pasivní diateze

V datové komponentě slovníku je možnost tvořit recipientní pasivní konstrukce vyznačena hodnotou atributu diat: recipient. Příznakové konstrukce této diateze se tvoří u všech sloves, která ji umožňují, na základě jediného **pravidla G17**, `recipient`:

Recipientní diateze Společné pravidlo	Pravidlo G17 `recipient`
podmínky	diat: recipient ACT_1 & X_3 [PAT|ADDR|BEN]
slovesná forma	→ *dostat*|*dostávat* + participium trpné, 4. pád[1-23]
shoda	participium trpné: číslo+rod, Y_4 [PAT|EFF] *dostat*|*dostávat*: číslo+rod+osoba, X
ACT	* → od+2, (7)
X	* → 1
obligatornost	X

Markéta Lopatková – Research Projects

Valency lexicon of Czech verbs – VALLEX

- complex predicates with light verbs
- Combining Words: Syntactic Properties of Czech Multiword Expressions with Light Verbs, GAČR 2015-17, PI Václava Kettnerová

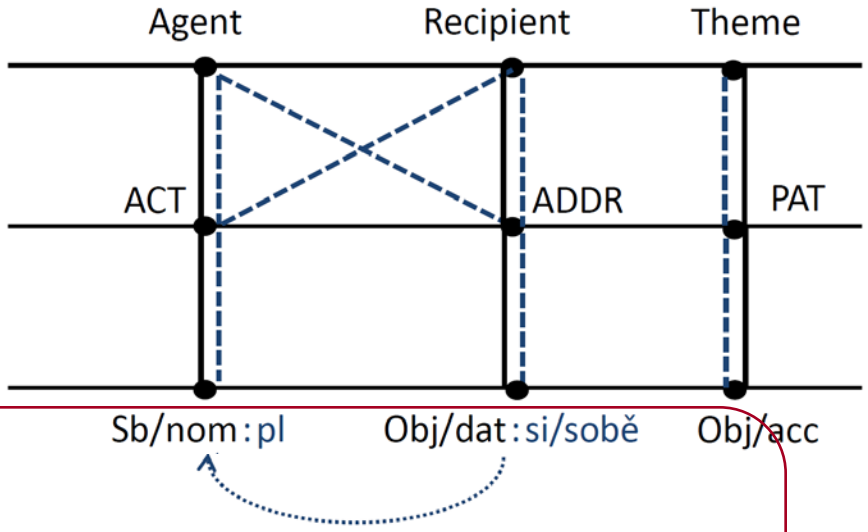
  - collocations of light verbs and predicative nouns (light verb constructions)
  - *two* syntactic elements function as *a single predicate:*

    light verbs ~ syntactic center of CPs
    predicative nouns ~ semantic center of CPs

Markéta Lopatková – Research Projects

Valency lexicon of Czech verbs – VALLEX

- GAČR project proposal:
- Between Reciprocity and Reflexivity: The Case of Czech Reciprocal Constructions



svěřovat 'entrust'

Jana svěřuje děti sestře Marii.
Jane$_{ACT.nom.sg}$ entrusts$_{pres.3sg}$ children$_{PAT.acc}$ sister Mary$_{ADD}$
'Jane entrusts her children to her sister Mary.'

Jana a Marie si vzájemně svěřují děti.
(Jane$_{nom.sg}$ and$_{conj}$ Mary$_{nom.sg}$)$_{ACT}$ REFL$_{ADDR.dat}$ to each other entrust$_{pres.3pl}$ children$_{PAT.acc}$
'Jane entrusts her children to Mary and at the same time Mary entrusts her children to Jane.'

Responsibilities of the Head of the Institute

Central funding

- PROVOZ … teaching money
  - salaries: ca 1.18 mil. CZK salaries (1.65 full contracts)
  - others: 603 th. CZK (traveling, …)
- PROGRES … research money (formerly PRVOUK)
  - salaries: ca 2.95 mil. (ca 5.5 full contracts)
  - other: 500 th. CZK (traveling, …)
- projects co-financing
  - GAČR … salaries: 711 th.
  - OP … salaries: 437 th.
    others: 632 th.
- Specific Research
  - scholarships: ca 240 th. CZK
  - other costs: 140 th (traveling, …)

Reporting and reporting and reporting

Markéta Lopatková – Teaching

Master program **Matematická lingvistika** (IML)
/ Computational Linguistics (IMLA)
("teacher responsible for the program")

Courses:

- Mathematical analysis
  winter + summer term, a practical course, BSc.
- Prague Dependency Treebank
  summer term, with Jiří Mírovský
- Mathematical Methods in Linguistics (??)

Supervising:

- 3 PhD students

Markéta Lopatková – Teaching

EM Language and Communication Technologies (LCT)

- ERASMUS MUNDUS double degree (together with Vláďa Kuboň)
- funded by EU: 2007-12, 2013-19
- 7 student for 2017-18:
  3+1 first year students
  3 second year students
  (plus 1+1 for 2018/19)

- EM LCT statistics (2007/08-2016/17):
  - enrolled in Prague: 43
  - graduated 33
  - delayed 2
  - failed 3
  - year 2 2+3

  plus 3 non-LCT master students

Markéta Lopatková – Others

- scientific board FF UK
- Prague Linguistic Cercle
- editorial board:

  Slovo a slovesnost
  Korpus – Gramatika – Axiologie
- coordinator of Erasmus exchange:

  Bolzano, Trento, Groningen, San Sebastian/Donostia
- member of program and organizing committees and reviewer

David Mareček

Research until now:

- HimL - experiments using Nematus - attention-based encoder-decoder NMT tool
- adding valency frames, functors, interleaved lemmas and tags

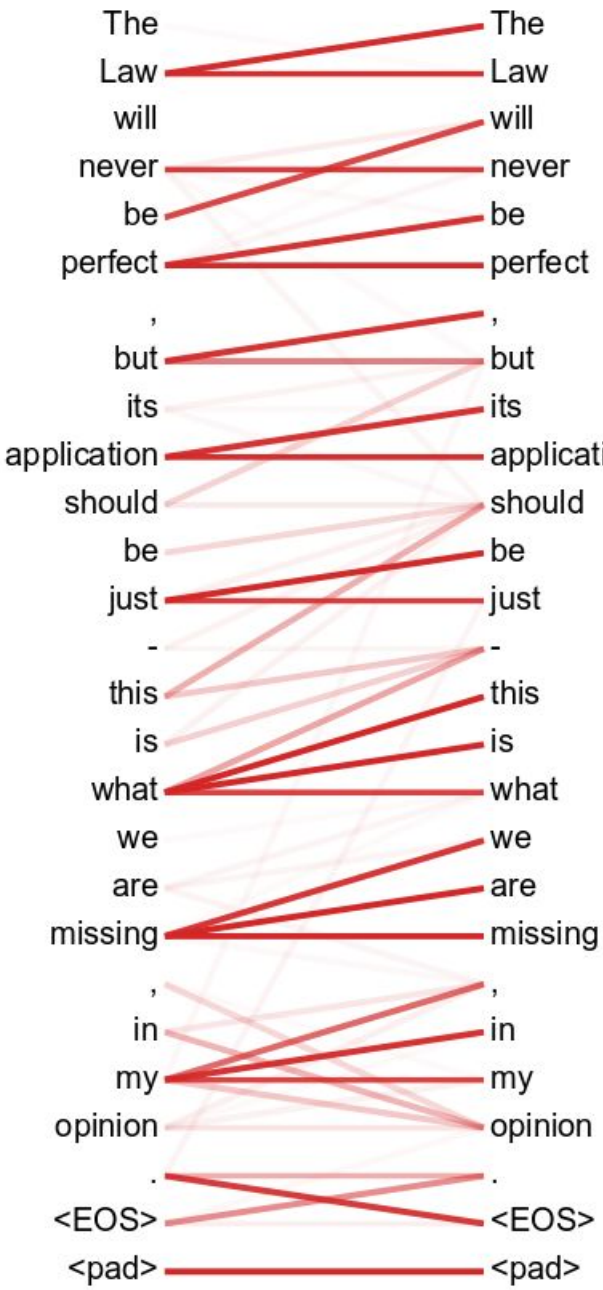
Teaching:

NPFL097 Selected Problems in Machine Learning

- Unsupervised machine learning, Bayesian inference, Gibbs sampling, ...

Would like to do:

- interpretability of neural networks
- analysis of (self-)attention in transformer and comparison with dependency trees


The
Law
will
never
be
perfect
,
but
its
application
should
be
just
-
this
is
what
we
are
missing
,
in
my
opinion
.
<EOS>
<pad>
The
Law
will
never
be
perfect
,
but
its
application
should
be
just
-
this
is
what
we
are
missing
,
in
my
opinion
.
<EOS>
<pad>


The
Law
will
never
be
perfect
,
but
its
application
should
be
just
-
this
is
what
we
are
missing
,
in
my
opinion
.
<EOS>
<pad>
The
Law
will
never
be
perfect
,
but
its
application
should
be
just
-
this
is
what
we
are
missing
,
in
my
opinion
.
<EOS>
<pad>

Personal Profile

Nikita Mediankin

ÚFAL MFF UK

14th Sep 2017, Sedlec-Prčice

Deep Syntactic Representation across Languages

Motivation

1. There are many independent incarnations of the same ideas for deep syntax.
2. Deep syntax is essentially a multilingual idea:
   - Abstraction from the grammar of the specific language.
   - Usually accompanied by a valency or functional lexicon of sorts.
   - Quite a few frameworks are in fact used or were developed for machine translation.
3. Now we have multilingual data with unified morphology and surface syntax because of the Universal Dependencies project.

Goals

- Let's try to decompose them and compare their components.
- We could use or not use certain ideas to create a deep syntactic representation for UDs...
- ...and test the actual applicability of created model on multilingual data.

Deep Syntactic Representation across Languages

First step: digging into existing Frameworks

- Functional Generative Decription (Tectogrammatical layer)
- Meaning—Text Theory (Deep Syntactic layer)
- PropBank Family (PropBank, NomBank, Penn Discourse Treebank, OntoNotes)
- Abstract Meaning Representation
- Microsoft Logical Forms
- Enhanced Universal Dependencies
- ...and 7 or 8 other.

Joint work with Magda Ševčíková, Dan Zeman, and Zdeněk Žabokrtský.

PoliSys Project: Summarization Task

Any Existing Czech summarization datasets?

- MultiLing Shared Task (http://multiling.iit.demokritos.gr):
  - part of a multilingual dataset;
  - 40 documents;
  - manually created from Czech Wikipedia articles.
- ...and not much else we could find.

SumeCzech

- News articles from novinky.cz, lidovky.cz, idnes.cz, denik.cz (ceskenoviny.cz coming soon).
- Obtained raw data from CommonCrawl project, cleaned up, extracted for each document:
  - headline (1 sentence);
  - summary (1-4 sentences);
  - full text.
- Currently approx. 550K documents.

PoliSys Project: Summarization Task

Three basic summarization setups

- full text → summary;
- full text → headline;
- summary → headline.

Experiments

- Unsupervised extractive baselines (first 1/3, TextRank, LexRank etc.).
- Tom Kocmi: NN-based abstractive summarization (summary → headline).

Evaluation

- ROUGE-raw: -1, -2, -L without preprocessing;
- ROUGE-cz-stems: -L with Czech stemming;
- ROUGE-cz-lemmas: -L with Czech lemmatization using MorphoDiTa.

I also did...

Python API for DeriNet

https://github.com/tiefling-cat/derinet-python

Marie Mikulová

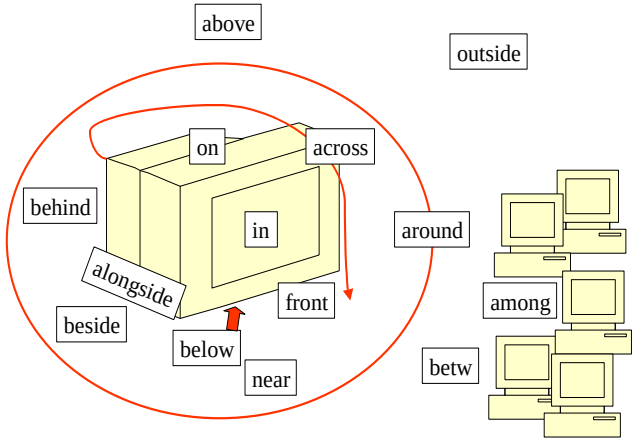
Prague Dependency Treebank Consolidated PDT-C 1.0

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

LDC 2020

text PDT
PDTSC speech
translation PCEDT
FAUST internet
Morphology
Syntax
Semantics

Subcategorization of Adverbial Meanings Based on Corpus Data



Marie Mikulová, Jarmila Panevová, Veronika Kolářová, Eduard Bejček
2019

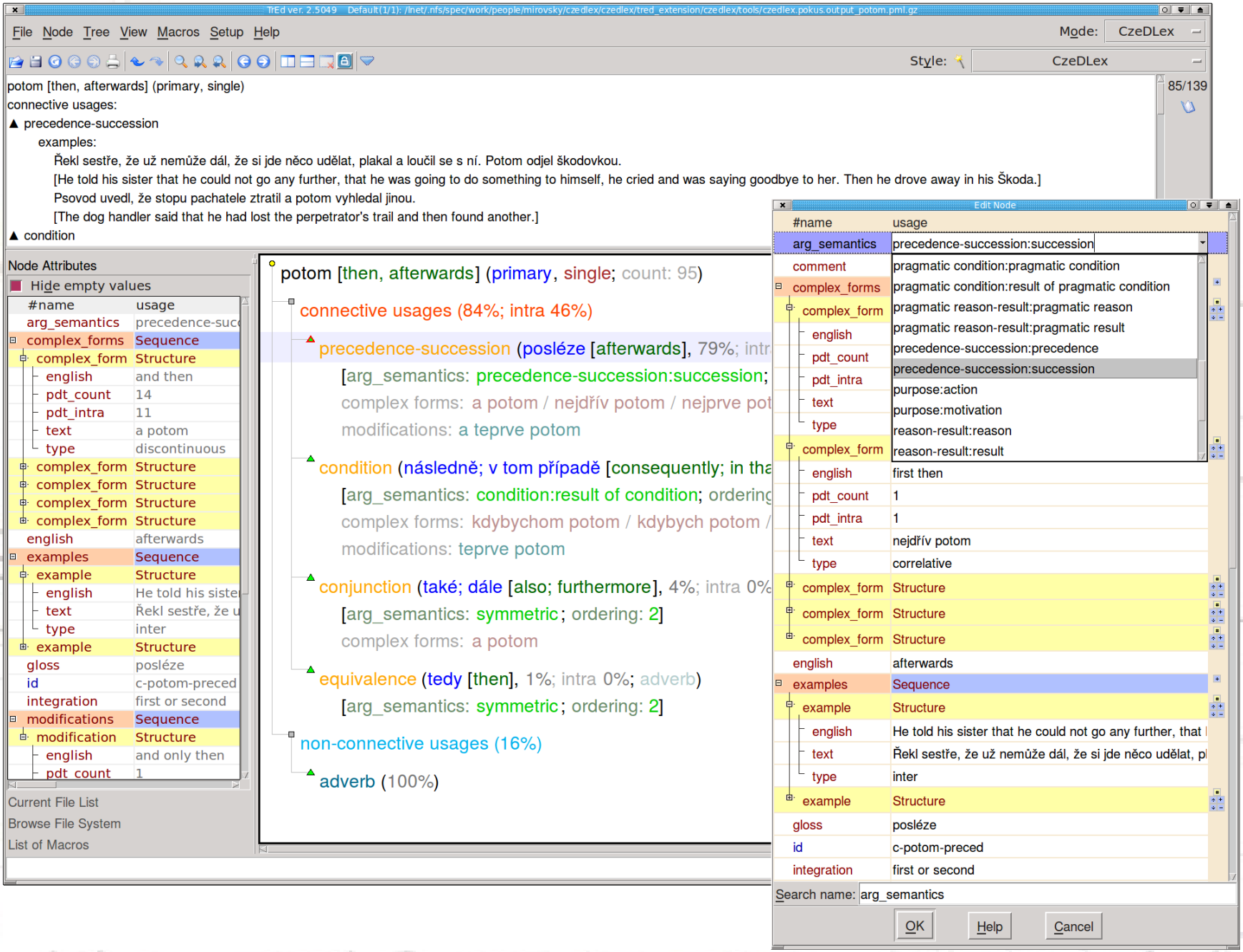
GAČR 2017-2019

Jiří Mírovský

Discourse-related activities

- maintaining the annotated **data** and **software** (PDiT 2.0)
- maintaining **TrEd extension for PDT 3.0** (and several others)
- working on NAKI II project – **measuring text coherence**
  - (using Treex & WEKA)
- Management Committee and Steering Committee member of European **project COST TextLink**
- project **COST-cz TextLink** – development of **CzeDLex** (Lexicon of Czech Discourse connectives)
  - (using PML and TrEd)

CzeDLex


File Node Tree View Macros Setup Help
Mode: CzeDLex
Style: CzeDLex
potom [then, afterwards] (primary, single)
connective usages:
▲ precedence-succession
examples:
Řekl sestře, že už nemůže dál, že si jde něco udělat, plakal a loučil se s ní. Potom odjel škodovkou.
[He told his sister that he could not go any further, that he was going to do something to himself, he cried and was saying goodbye to her. Then he drove away in his Škoda.]
Psovod uvedl, že stopu pachatele ztratil a potom vyhledal jinou.
[The dog handler said that he had lost the perpetrator's trail and then found another.]
▲ condition
85/139
Node Attributes
Hide empty values
#name usage
arg_semantics precedence-succ
complex_forms Sequence
complex_form Structure
english and then
pdt_count 14
pdt_intra 11
text a potom
type discontinuous
complex_form Structure
complex_form Structure
complex_form Structure
complex_form Structure
english afterwards
examples Sequence
example Structure
english He told his siste
text Řekl sestře, že u
type inter
example Structure
gloss posléze
id c-potom-preced
integration first or second
modifications Sequence
modification Structure
english and only then
pdt_count 1
Current File List
Browse File System
List of Macros
potom [then, afterwards] (primary, single; count: 95)
connective usages (84%; intra 46%)
precedence-succession (posléze [afterwards], 79%; intr
[arg_semantics: precedence-succession:succession;
complex forms: a potom / nejdřív potom / nejprve pot
modifications: a teprve potom
condition (následně; v tom případě [consequently; in tha
[arg_semantics: condition:result of condition; ordering
complex forms: kdybychom potom / kdybych potom /
modifications: teprve potom
conjunction (také; dále [also; furthermore], 4%; intra 0%
[arg_semantics: symmetric; ordering: 2]
complex forms: a potom
equivalence (tedy [then], 1%; intra 0%; adverb)
[arg_semantics: symmetric; ordering: 2]
non-connective usages (16%)
adverb (100%)
Edit Node
#name usage
arg_semantics precedence-succession:succession
comment
complex_forms
complex_form
english
pdt_count
pdt_intra
text
type
complex_form
english first then
pdt_count 1
pdt_intra 1
text nejdřív potom
type correlative
complex_form Structure
complex_form Structure
complex_form Structure
english afterwards
examples Sequence
example Structure
english He told his sister that he could not go any further, that
text Řekl sestře, že už nemůže dál, že si jde něco udělat, p
type inter
example Structure
gloss posléze
id c-potom-preced
integration first or second
pragmatic condition:pragmatic condition
pragmatic condition:result of pragmatic condition
pragmatic reason-result:pragmatic reason
pragmatic reason-result:pragmatic result
precedence-succession:precedence
precedence-succession:succession
purpose:action
purpose:motivation
reason-result:reason
reason-result:result
Search name: arg_semantics
OK
Help
Cancel

Jiří Mírovský

ÚFAL-wide activities

- ordering/maintaining **software from LDC** (and other sw, e.g. dictionaries, Adobe Acrobat, ...), plus associated wiki web pages
- maintaining the **Amoeba** database for ÚFAL (with V. Kuboň+)
- maintaining web pages with **PML-TQ documentation** and **examples**
- **searching in PML-TQ** on request
- maintaining **PML-TQ search servers** for PDT 3.0, PDiT 2.0, ...
- maintaining ÚFAL **web pages** for PDiT 2.0, PDT 3.0 (and a couple of others)
- preparing the publication of **PDTSC [12].0** (with M. Mikulová)
- teaching: **practical sessions** for Markéta's lectures about PDT (NPFL075)

Tomáš Musil

- starting PhD this year
- research interests
  - AI
  - machine learning
  - neural networks
    - neural machine translation
    - Neural Monkey
  - (analytical) philosophy (of language)
- dissertation
  - Exploring Language Principles with Respect to Algorithms of Deep Neural Networks
    - what is the essence of language?
    - can we learn something about it from deep learning?
  - supervisor: David Mareček

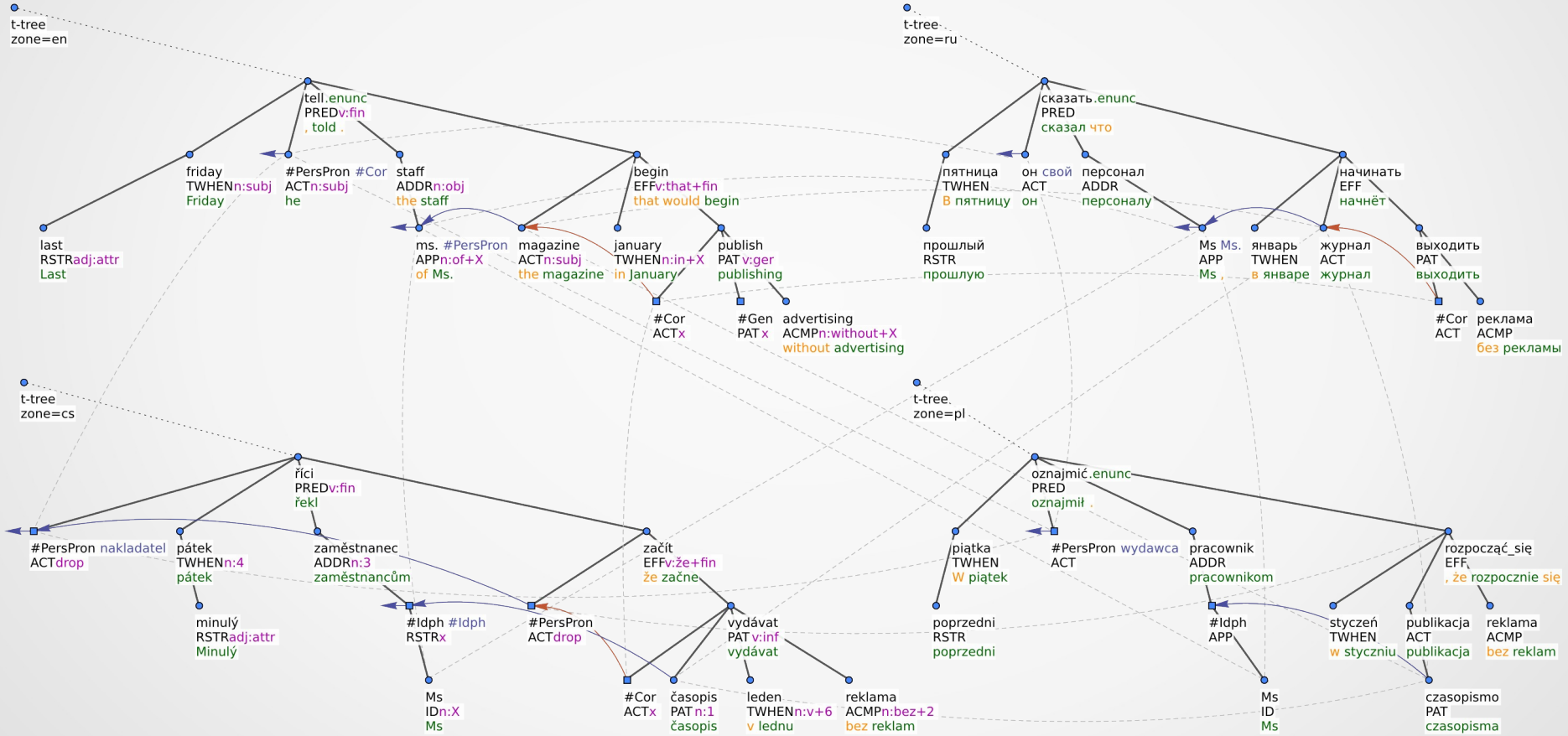
Michal Novák

- **GAUK: Cross-lingual approaches to coreference resolution**
  - Coreference Resolution (Treex CR)
  - cross-lingual CR
  - semi-supervised approaches for cross-lingual CR
  - machine-learning: VowpalWabbit, MLyn (https://github.com/michnov/MLyn)
  - the central part of my upcoming PhD thesis
- **GAČR: Structure of coreferential chains in parallel language data**
  - with Anja Nedoluzhko
  - comparison of languages in terms of how they express coreference
  - coreference projection in parallel data
  - AnaphBus vs. PAWS (Parallel Anaphoric WSJ)
    - with Anja and Maciej Ogrodniczuk (Polish Academy of Sciences)
    - 1k sent quartets in English, Czech, Russian and Polish from WSJ
    - coreference in tecto-like style

AnaphBus vs. 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.

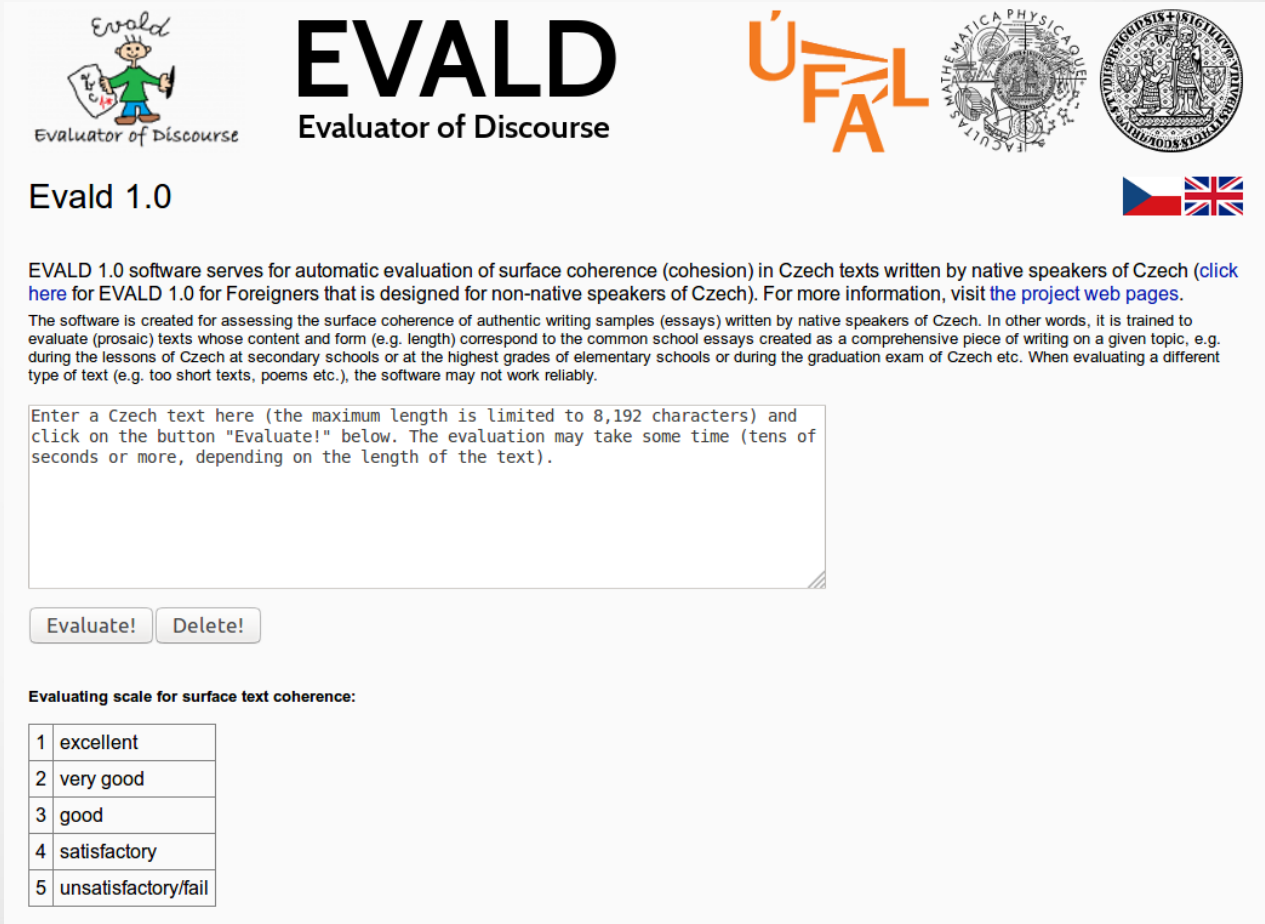


CS: Minulý pátek řekl zaměstnancům Ms., že časopis v lednu začne vydávat bez reklam.

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

Michal Novák

- **NAKI: EVALD (Evaluator of Discourse)**
  - with Kačka and Majda Rýsová, Jirka Mírovský, prof. Hajičová
  - assessing the level of coherence in students' essays
  - Treex, Docker



Michal Novák

- **ÚFAL Beer Committee Founding Member**
  - the last Beer was yesterday (if you do not remember)
  - the next Beer is on October 12th
- **ÚFAL's Publishing House**
  - supplying Karolinum bookstore with books published at ÚFAL
  - offering the books at events organized by ÚFAL
  - administration of the related web pages (http://ufal.cz/books)

http://ufal.cz/books

ÚFAL's Publishing House
Annual report

Sales and donations of ÚFAL books

Book	Sales		Donations		Other		Total	
	2016/17	All years	2016/17	All years	2016/17	All years	2016/17	All years
Ondřej Bojar: Exploiting linguistic data in MT	1	**7**	16	**35**	5	**5**	22	**47**
Petr Homola: Syntatic analysis in MT	1	**5**	14	**30**	6	**6**	21	**41**
Pavel Pecina: Lexical association measures	1	**9**	14	**26**	4	**4**	19	**39**
Ondřej Bojar: Čeština a strojový překlad	5	**20**	4	**6**	2	**2**	11	**28**
Silvie Cinková: Words that Matter		**2**	5	**11**	10	**10**	15	**23**
Jiří Mírovský: Searching in the PDT		**3**	4	**8**	3	**3**	7	**14**
Radek Čech: Tematická koncentrace textu	1	**5**	1	**2**	3	**3**	5	**10**
Barbora Štěpánková: Aktualizátory ve výstavbě textu	1	**7**	2	**2**		**0**	3	**9**
Anna Nědolužko: Rozšířená textová koreference		**5**	3	**3**		**0**	3	**8**
Kateřina Rysová: O slovosledu		**5**	1	**2**		**0**	1	**7**
Marie Mikulová: Významová reprezentace elipsy		**4**	1	**1**		**0**	1	**5**
Zdeňka Urešová: Valence sloves v PDT		**3**	2	**2**		**0**	2	**5**
Zdeňka Urešová: Valenční slovník PDT-Vallex		**3**	2	**2**		**0**	2	**5**
Magda Ševčíková: Funkce kondicionálu		**2**	1	**1**		**0**	1	**3**
Zikánová et al.: Discourse and Coherence		**1**		**0**	1	**1**	1	**2**
Total	**10**	**81**	**70**	**131**	**34**	**34**	**114**	**246**

Sales and donations of Ú

- taken by the author
- taken by passersby
- moved to another place without letting me know
- my mistake
- mystery

Book	Sales		Donations		Other		Total	
	2016/17	All years	2016/17	All years	2016/17	All years	2016/17	All years
Ondřej Bojar: Exploiting linguistic data in MT	1	**7**	16	**35**	5	**5**	22	**47**
Petr Homola: Syntatic analysis in MT	1	**5**	14	**30**	6	**6**	21	**41**
Pavel Pecina: Lexical association measures	1	**9**	14	**26**	4	**4**	19	**39**
Ondřej Bojar: Čeština a strojový překlad	5	**20**	4	**6**	2	**2**	11	**28**
Silvie Cinková: Words that Matter		**2**	5	**11**	10	**10**	15	**23**
Jiří Mírovský: Searching in the PDT		**3**	4	**8**	3	**3**	7	**14**
Radek Čech: Tematická koncentrace textu	1	**5**	1	**2**	3	**3**	5	**10**
Barbora Štěpánková: Aktualizátory ve výstavbě textu	1	**7**	2	**2**		**0**	3	**9**
Anna Nědolužko: Rozšířená textová koreference		**5**	3	**3**		**0**	3	**8**
Kateřina Rysová: O slovosledu		**5**	1	**2**		**0**	1	**7**
Marie Mikulová: Významová reprezentace elipsy		**4**	1	**1**		**0**	1	**5**
Zdeňka Urešová: Valence sloves v PDT		**3**	2	**2**		**0**	2	**5**
Zdeňka Urešová: Valenční slovník PDT-Vallex		**3**	2	**2**		**0**	2	**5**
Magda Ševčíková: Funkce kondicionálu		**2**	1	**1**		**0**	1	**3**
Zikánová et al.: Discourse and Coherence		**1**		**0**	1	**1**	1	**2**
Total	**10**	**81**	**70**	**131**	**34**	**34**	**114**	**246**

Sales and donations of ÚFAL books

Book	Sales		Donations		Other		Total	
	2016/17	All years	2016/17	All years	2016/17	All years	2016/17	All years
Ondřej Bojar: Exploiting linguistic data in MT	1	**7**	16	**35**	5	**5**	22	**47**
Petr Homola: Syntatic analysis in MT	1	**5**	14	**30**	6	**6**	21	**41**
Pavel Pecina: Lexical association measures	1	**9**	14	**26**	4	**4**	19	**39**
Ondřej Bojar: Čeština a strojový překlad	5	**20**	4	**6**	2	**2**	11	**28**
Silvie Cinková: Words that Matter		**2**	5	**11**	10	**10**	15	**23**
Jiří Mírovský: Searching in the PDT		**3**	4	**8**	3	**3**	7	**14**
Radek Čech: Tematická koncentrace textu	1	**5**	1	**2**	3	**3**	5	**10**
Barbora Štěpánková: Aktualizátory ve výstavbě textu	1	**7**	2	**2**		**0**	3	**9**
Anna Nědolužko: Rozšířená textová koreference		**5**	3	**3**		**0**	3	**8**
Kateřina Rysová: O slovosledu		**5**	1	**2**		**0**	1	**7**
Marie Mikulová: Významová reprezentace elipsy		**4**	1	**1**		**0**	1	**5**
Zdeňka Urešová: Valence sloves v PDT		**3**	2	**2**		**0**	2	**5**
Zdeňka Urešová: Valenční slovník PDT-Vallex		**3**	2	**2**		**0**	2	**5**
Magda Ševčíková: Funkce kondicionálu		**2**	1	**1**		**0**	1	**3**
Zikánová et al.: Discourse and Coherence		**1**		**0**	1	**1**	1	**2**
Total	**10**	**81**	**70**	**131**	**34**	**34**	**114**	**246**

- change in sales: -42%
- change in donations: +60%

Sales and donations of ÚFAL books

Book	Sales		Donations		Other		Total	
	2016/17	All years	2016/17	All years	2016/17	All years	2016/17	All years
Ondřej Bojar: Exploiting linguistic data in MT	1	**7**	16	**35**	5	**5**	22	**47**
Petr Homola: Syntatic analysis in MT	1	**5**	14	**30**	6	**6**	21	**41**
Pavel Pecina: Lexical association measures	1	**9**	14	**26**	4	**4**	19	**39**
Ondřej Bojar: Čeština a strojový překlad	5	**20**	4	**6**	2	**2**	11	**28**
Silvie Cinková: Words that Matter		**2**	5	**11**	10	**10**	15	**23**
Jiří Mírovský: Searching in the PDT		**3**	4	**8**	3	**3**	7	**14**
Radek Čech: Tematická koncentrace textu	1	**5**	1	**2**	3	**3**	5	**10**
Barbora Štěpánková: Aktualizátory ve výstavbě textu	1	**7**	2	**2**		**0**	3	**9**
Anna Nědolužko: Rozšířená textová koreference		**5**	3	**3**		**0**	3	**8**
Kateřina Rysová: O slovosledu		**5**	1	**2**		**0**	1	**7**
Marie Mikulová: Významová reprezentace elipsy		**4**	1	**1**		**0**	1	**5**
Zdeňka Urešová: Valence sloves v PDT		**3**	2	**2**		**0**	2	**5**
Zdeňka Urešová: Valenční slovník PDT-Vallex		**3**	2	**2**		**0**	2	**5**
Magda Ševčíková: Funkce kondicionálu		**2**	1	**1**		**0**	1	**3**
Zikánová et al.: Discourse and Coherence		**1**		**0**	1	**1**	1	**2**
Total	**10**	**81**	**70**	**131**	**34**	**34**	**114**	**246**

- change in sales: -42%
- change in donations: +60%

No new publications

Sales and donations of ÚFAL books

Book	Sales		Donations		Other		Total	
	2016/17	All years	2016/17	All years	2016/17	All years	2016/17	All years
Ondřej Bojar: Exploiting linguistic data in MT	1	**7**	16	**35**	5	**5**	22	**47**
Petr Homola: Syntatic analysis in MT	1	**5**	14	**30**	6	**6**	21	**41**
Pavel Pecina: Lexical association measures	1	**9**	14	**26**	4	**4**	19	**39**
Ondřej Bojar: Čeština a strojový překlad	5	**20**	4	**6**	2	**2**	11	**28**
Silvie Cinková: Words that Matter		**2**	5	**11**	10	**10**	15	**23**
Jiří Mírovský: Searching in the PDT		**3**	4	**8**	3	**3**	7	**14**
Radek Čech: Tematická koncentrace textu	1	**5**	1	**2**	3	**3**	5	**10**
Barbora Štěpánková: Aktualizátory ve výstavbě textu	1	**7**	2	**2**		**0**	3	**9**
Anna Nědolužko: Rozšířená textová koreference		**5**	3	**3**		**0**	3	**8**
Kateřina Rysová: O slovosledu		**5**	1	**2**		**0**	1	**7**
Marie Mikulová: Významová reprezentace elipsy		**4**	1	**1**		**0**	1	**5**
Zdeňka Urešová: Valence sloves v PDT		**3**	2	**2**		**0**	2	**5**
Zdeňka Urešová: Valenční slovník PDT-Vallex		**3**	2	**2**		**0**	2	**5**
Magda Ševčíková: Funkce kondicionálu		**2**	1	**1**		**0**	1	**3**
Zikánová et al.: Discourse and Coherence		**1**		**0**	1	**1**	1	**2**
Total	**10**	**81**	**70**	**131**	**34**	**34**	**114**	**246**

- change in sales: -42%
- change in donations: +60%

No new publications

Many events:
DRMC 2016 (KONTAKT II)
TextLink Training School 2017
EAMT 2017
Tyden diverzity FF UK
TSD 2017

How to increase the distribution?

Book	In stock	Expected years
Kateřina Rysová: O slovosledu	0	0
Zikánová et al.: Discourse and Coherence	0	0
Pavel Pecina: Lexical association measures	15	2
Ondřej Bojar: Exploiting linguistic data in MT	26	3
Barbora Štěpánková: Aktualizátory ve výstavbě textu	14	4
Petr Homola: Syntatic analysis in MT	58	8
Ondřej Bojar: Čeština a strojový překlad	47	8
Radek Čech: Tematická koncentrace textu	65	11
Silvie Cinková: Words that Matter	33	12
Jiří Mírovský: Searching in the PDT	61	> 15
Anna Nědolužko: Rozšířená textová koreference	65	> 15
Zdeňka Urešová: Valence sloves v PDT	47	> 15
Zdeňka Urešová: Valenční slovník PDT-Vallex	67	> 15
Marie Mikulová: Významová reprezentace elipsy	99	> 15
Magda Ševčíková: Funkce kondicionálu	82	> 15
Total	679	

How to increase the distribution?

Book	In stock	Expected years
Kateřina Rysová: O slovosledu	0	0
Zikánová et al.: Discourse and Coherence	0	0
Pavel Pecina: Lexical association measures	15	2
Ondřej Bojar: Exploiting linguistic data in MT	26	3
Barbora Štěpánková: Aktualizátory ve výstavbě textu	14	4
Petr Homola: Syntatic analysis in MT	58	8
Ondřej Bojar: Čeština a strojový překlad	47	8
Radek Čech: Tematická koncentrace textu	65	11
Silvie Cinková: Words that Matter	33	12
Jiří Mírovský: Searching in the PDT	61	> 15
Anna Nědolužko: Rozšířená textová koreference	65	> 15
Zdeňka Urešová: Valence sloves v PDT	47	> 15
Zdeňka Urešová: Valenční slovník PDT-Vallex	67	> 15
Marie Mikulová: Významová reprezentace elipsy	99	> 15
Magda Ševčíková: Funkce kondicionálu	82	> 15
Total	679	

- Suggestions for the authors:
  - Take care of your book's distribution
  - Conferences, workshops, meetings
- Suggestions for the others:
  - Let me know if you organize an event or you know about an event, where we can offer books
    - ITAT / SloNLP 2017

Books are rather for ...

Books are rather for ...

than for ...

Pavel Pecina

- PI:
  - **H2020 KConnect** (2015-17) – medical text MT
  - **GAČR CEMI** (2012-18) – multimodal data interpretation
- Teaching:
  - **NPFL067/8** (with prof. Hajič) - Statistical NLP
  - **NPFL103** - Information Retrieval
  - **B4M36NL** (FEL ČVUT)– Intro to NLP
- Students:
  - **Petra Galuščáková** - speech segmentation and retrieval
  - **Shadi Saleh** - cross-lingual information retrieval
  - **Jindřich Libovický** - reading text in images
  - **Jan Hajič jr.** - optical music recognition
  - **Michal Auersperger** - document embeddings
  - **Karolína Burešová** - text simplification

Martin Popel

- NLP frameworks: Treex, **Udapi** http://udapi.github.io
  - Perl, Java, **Python** see our paper about Udapi
  - 100 time faster than Treex
  - native support for Universal Dependencies (CoNLL-U)
  - tree visualizations, querying, exports, parsing (UDPipe)
- **Universal Dependencies** (CoNLL 2017), Dan's GAČR **Manyla**
- **TectoMT** tectogrammatical machine translation
  - EN↔CS, EN↔ES, EN↔NL, EN↔PT, EN↔EU, Vowpal Wabbit
- **MT-ComparEval** (+Ondřej Klejch)

http://mt-compareval.ufal.cz upload your MT outputs

http://wmt.ufal.cz compare WMT17 systems

Martin Popel

- **PBML** (next deadline: January 12th 2018) + Dušan Variš
- **Technical reports** (2017 deadline: December 1st)
- **Teaching** autumn: Modern Methods in CL I ("Reading group")

  spring: Language Data Resources (+ZŽ)

  October: Natural language processing on computational cluster (+RR)

  introduction to ÚFAL for new PhD students
- My recent work: **Neural MT with Transformer and Tensor2tensor**

  state-of-the-art MT from Google Brain, fully open source

  better and faster than (deep) Nematus +6 BLEU (+4 BLEU)

  future plans: exploit syntax (multitask MT+parsing or src features)

  visualize and analyze self-attention (cf. dep. trees)

Mgr. Rudolf Rosa (rosa@ufal)

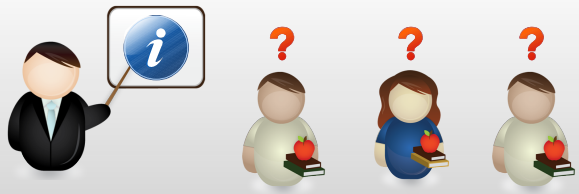
- **cross-lingual transfer** of dependency parsers (PhD, 4 years)
  - e.g. train a parser on Latvian → use it to parse Lithuanian
- small fun projects: simple chatbot, Czechizator...
- past: TectoMT&Depfix, HamleDT&UD, internship@Google

- **NPFL092**[ZŽ] Technology for NLP (Bash, **Python**, make, **svn/git**) **???**

NEW! **NPFL118**[MP] Natural language processing on computational cluster (aka intro for PhDs to using computers at ÚFAL)

NEW! **NPFL120**[DZ] Multilingual Natural Language Processing

- organizing **SloNLP** (Slovakoczech NLP workshop) ITAT
  - we welcome students & early-stage researchers!
- ÚFAL student ambassador



Kateřina Rysová

Projects:

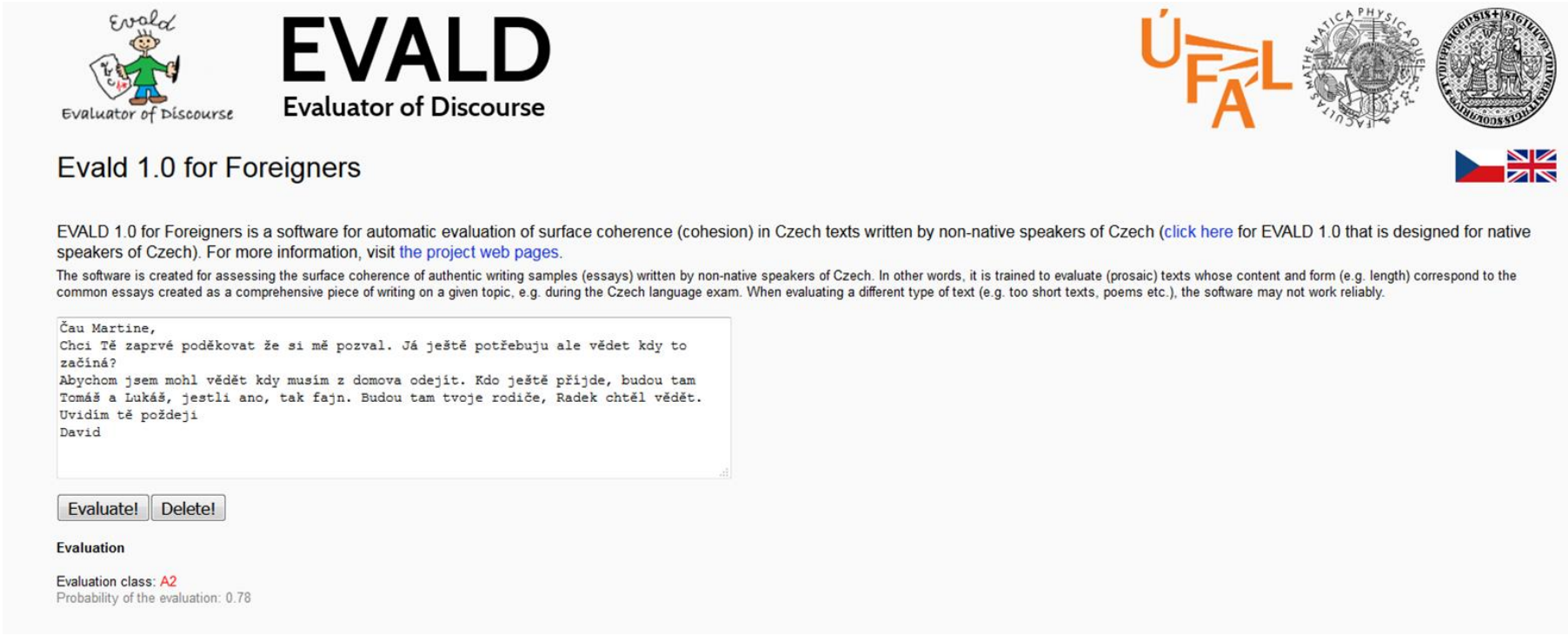
1) NAKI II: **EVALD – Evaluator of Discourse**

- 2016–2019
- classifier of texts written by **non-native speakers** of Czech (6 categories: from beginners to almost native speakers) and by **native speakers** of Czech (5 categories: school marks)

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

EVALD – Evaluator of Discourse

- available also online: https://lindat.mff.cuni.cz/services/evald-foreign/
- EVALD will be introduced at ÚFAL Monday seminar: 9th October 2017



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

- 2017–2019
- linguistically oriented discourse project
- delimitation and description of discourse connectives in Czech and German

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

Magdaléna Rysová

Involved in projects:

1) COST-cz – **TextLink: Structuring Discourse in Multilingual Europe** (2015–2017); PI: Jiří Mírovský

2) NAKI II – **Automatic Evaluation of Text Coherence in Czech** (2016–2019); PI: Kateřina Rysová

3) GAČR – **Anaphoricity of Connectives: Lexical Description and Billingual Corpus Analysis** (2017–2019); PI: Kateřina Rysová

4) COST – **Structuring Discourse in Multilingual Europe (TextLink)** (2014–2018); Czech PI: Jiří Mírovský

COST-cz

- Building a lexicon of Czech discourse connectives
- Entries for both primary (*proto*) and secondary connectives (*kvůli tomu*; *z tohoto důvodu*)

NAKI II

- Software applications (called EVALD – Evaluator of Discourse) for automatic evaluation of coherence in Czech texts written by 1) native and 2) non-native speakers of Czech
- Preparing datasets: finding and manually evaluating texts; finding linguistic features in which the individual classes differ (three fields: discourse, coreference and sentence information structure)

GAČR

- A comparative analysis of Czech and German cohesive means, especially of anaphoric connectives
- 2018: monograph – PhD thesis (defended in 2015: Discourse Connectives in Czech: From Centre to Periphery) enriched by research on anaphoricity of connectives

Magda Ševčíková

- PI of the projects
  - GA16-18177S *An Integrated Approach to Derivational and Inflectional Morphology of Czech*, 2016–2018
    - derivation of Czech, DeriNet database
  - Mobility France 7AMB16FR048 *Kontrastivní pohled na moderní českou morfologii s ohledem na frankofonní mluvčí*, 2016–2017
- PhD student Adéla Kalužová
- teaching 2017/18
  - NPFL006 *Introduction to Formal Linguistics*
    - winter term
  - NPFL121 *Selected topics from the Czech grammar*
    - with Anja Nedoluzhko and Šárka Zikánová, winter term
  - NPOZ009 *Professional language and style*
    - with Marie Mikulová, summer term
  - *Modern linguistic descriptions of English*
    - course on selected syntactic theories, master students of English philology, Faculty of Arts, winter term

DeriNet database

- Zdeněk Žabokrtský, Jonáš Vidra, Adéla Limburská, Vojtěch Hudeček; Nikita Mediankin, Milan Straka
- lexical database of Czech words (from MorfFlex CZ; nodes) connected with links corresponding to derivational relations (edges)
  - a word is linked to a word which it is supposed to be derived from
  - $učit > učitel > učitelka$
- 1,012K lemmas connected with 774K links in DeriNet 1.4
  - incl. 23K+ new derivational links between verbs (Adéla Kalužová)
  - 238K words not connected
- `http://ufal.mff.cuni.cz/derinet`
  - DeriNet Search `http://ufal.mff.cuni.cz/derinet/search`
  - DeriNet Viewer `http://ufal.mff.cuni.cz/derinet/viewer`

Derivation in Czech

- vowel and consonant alternations
- aspect as infl. feature expressed by derivation – Prof. J. Panevová
- aspect in action nouns
  - *výběr – vybrat / vybírat*
- derivational networks for (un/related) languages – M. Lango
- bound bases
  - *po-škodit* but *po-škozovat*: *škodit* > *poškodit* > *poškozovat*
  - *na-bídnout* and *na-bízet*
- modelling derivation of foreign words, e.g. *-ismus*
  - *socialismus* > *socialistický* but *fotbalismus* < *fotbalistický*
- compounds – A. Kalužová
- terminology
  - derivational morphology
    - Czech ling.: morphology=inflection vs word formation
  - borrowings and neoclassical formations
    - Czech ling.: cizí slovo, přejaté slov, výpůjčka, kalk, anglicismus, ...

Zdeňka Urešová

Contextually-based synonymy and valency of verbs in a bilingual setting

Kontextová synonymie a valence sloves v bilingvním prostředí

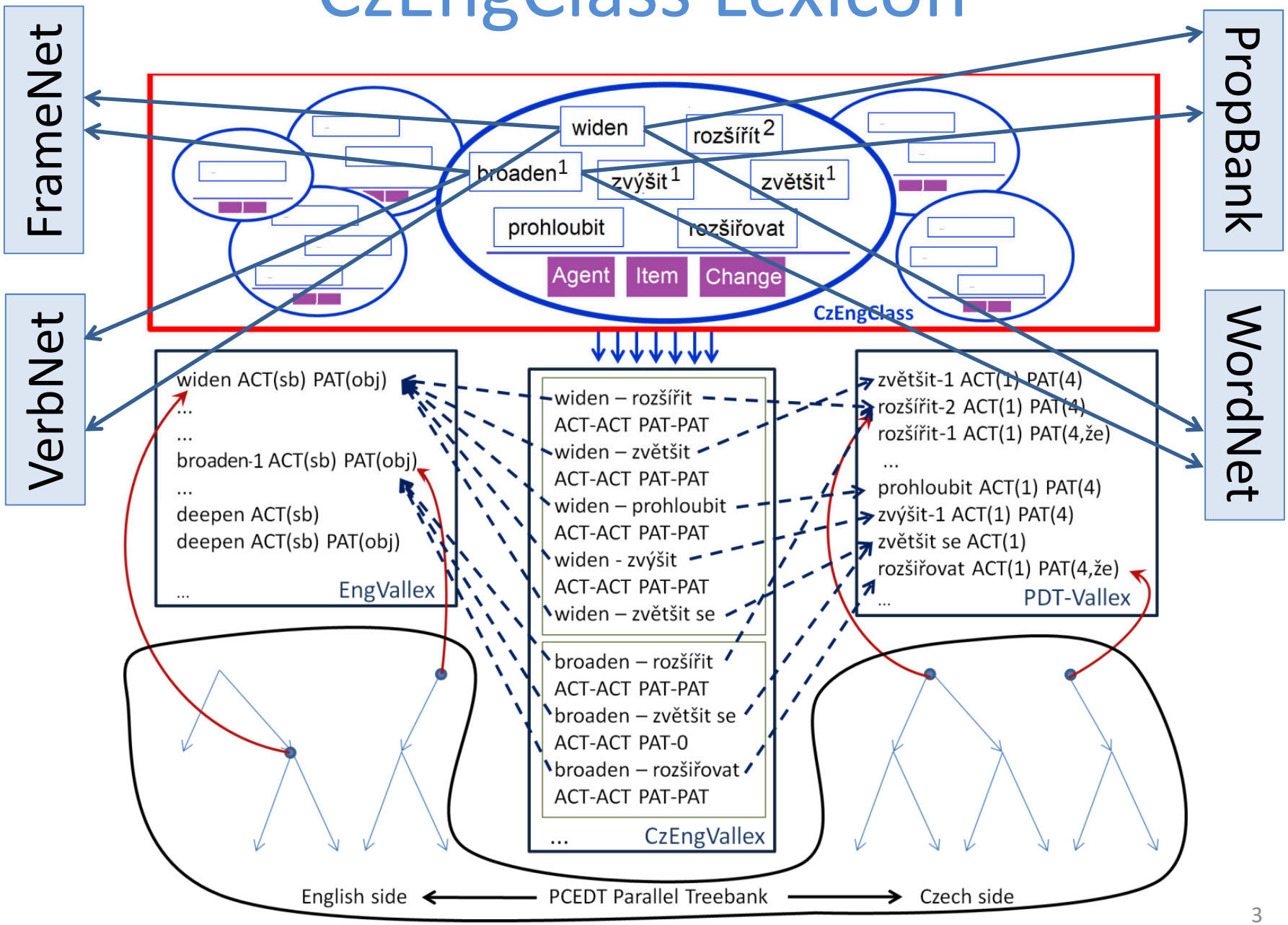
GAČR standard project (2017 – 2019)

- 3 people – Z. Urešová, E. Fučíková, E. Hajičová
- Theme:
  - verbal synonymy in translation (bilingual context, Czech-English)
    - based on the FGD (valency) theory
    - to explore semantic 'equivalence' of verb senses of different verbal lexemes
  - focus on valency behavior and semantic roles
  - assumption: bilingual context (translation) enables
    - to delimit synonymous verbs and so
    - to specify verb senses more precisely than monolingual text

Overview

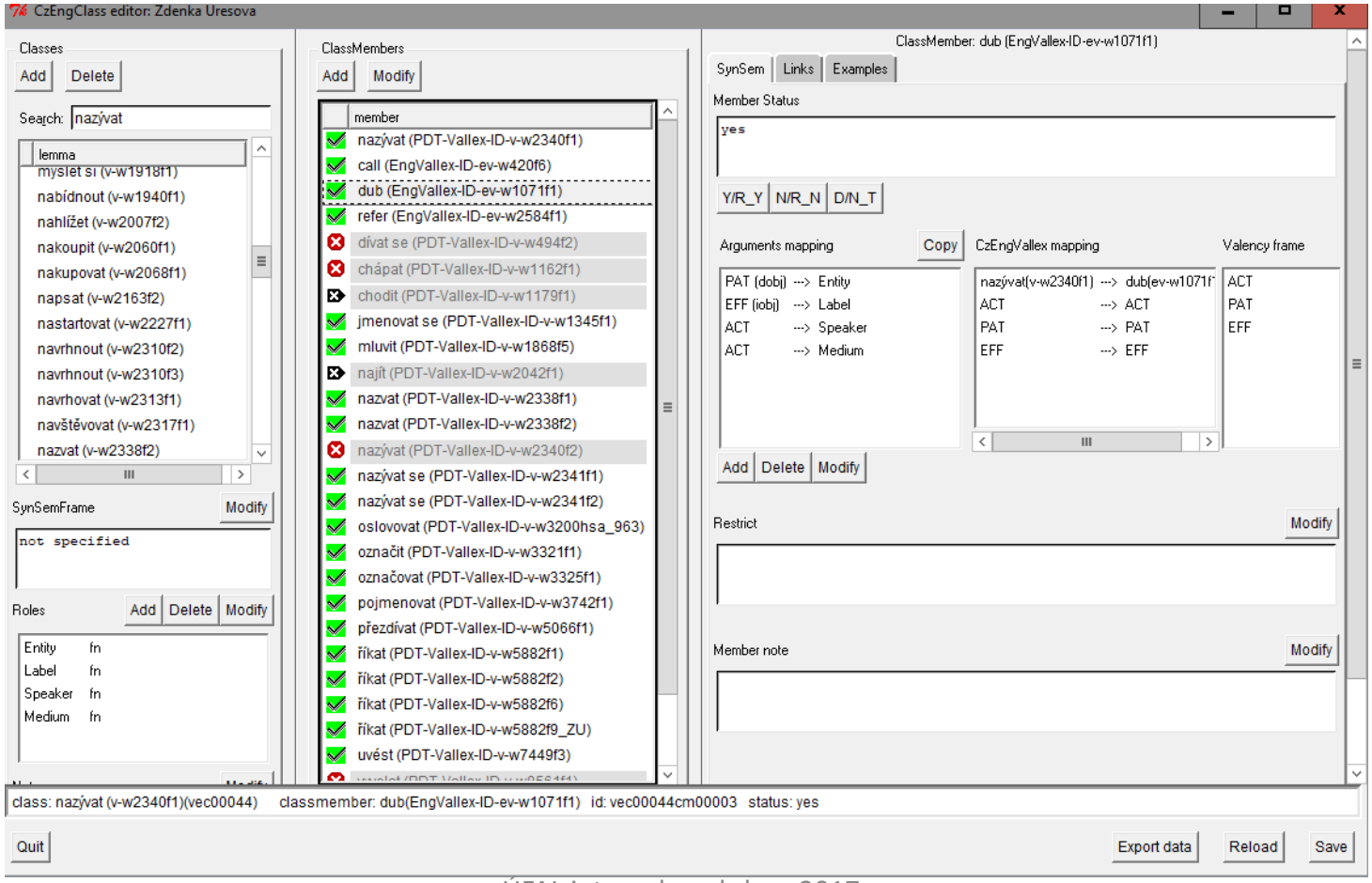
- Goal
  - to group verbs used as synonyms in Czech and English into (cross-lingual) synonym classes
- Approach: “bottom-up”, starting with evidence in bilingual corpus (vs. “topdown”, with predefined set of semantic or top-level synonym classes)
- Lexical Resources
  - Prague Dependency Treebank-style valency lexicons (PDT-Vallex, EngVallex and CzEngVallex)
  - Other (FrameNet, VerbNet, PropBank, Czech and English WordNets)
- Corpus Resources
  - The Prague Czech-English Dependency Treebank (PCEDT)
  - (Large monolingual corpora)
- Result
  - CzEngClass: lexicon of verb synonyms with valency mapped to semantic roles and linked to existing lexical resources

CzEngClass Lexicon



Eva Fučíková

- Technical support for the CzEngClass Lexicon Project
  - Data preparation
  - Annotation Editor



Dušan Variš

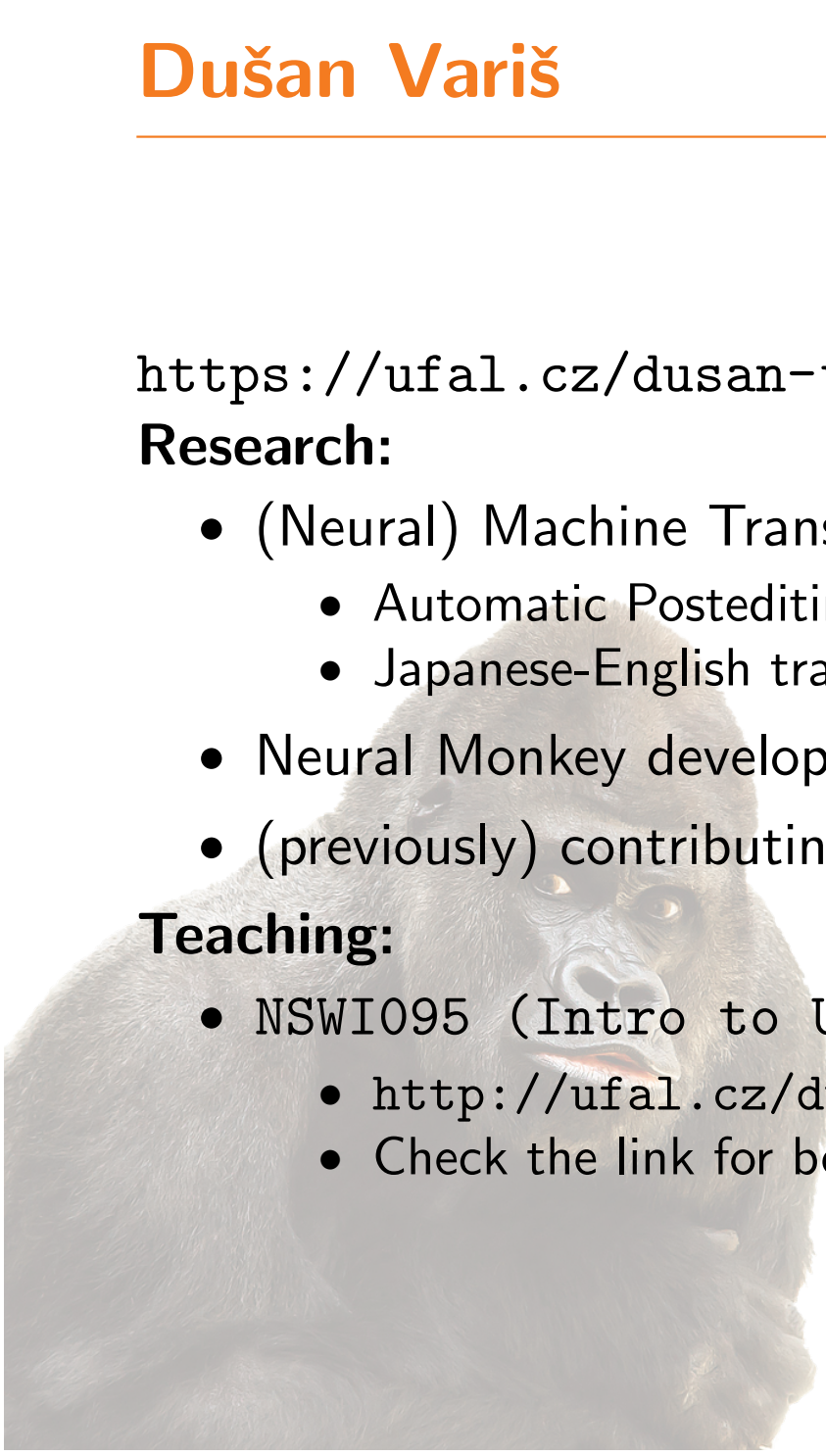
https://ufal.cz/dusan-varis

Research:

- (Neural) Machine Translation
  - Automatic Postediting of MT outputs
  - Japanese-English translation
- Neural Monkey development
- (previously) contributing to Treex

Teaching:

- NSWI095 (Intro to Unix)
  - http://ufal.cz/dusan-varis/nswi095
  - Check the link for beginner-level exercises ☺



Anna Vernerová

- KonText
  - inclusion of new corpora
  - help with using KonText and/or pml-tq

- NomVallex
  - noun valency
  - lexicon creation (no corpus annotation)
  - technical support

Katka Veselovská

- finishing a book on sentiment analysis

- GAČR: *On Linguistic Structure of Evaluative Meaning in Czech*
  - till 2017
  - from linguistic aspects to neural networks

- Next steps? Psycholinguistic experiments, multimodal data…?

Katka Veselovská

- Or: text analytics in forensic investigations
- Expertise: Semantic data science lead, forensic team at Deloitte + cooperation with Institute of Criminal Science (completely new pipeline for automatic text processing)
- i.e. forensic linguistics = sentiment + information extraction, author detection, coding speech detection, law language, suicide letters, plagiarism, threat communication, extremism in social media…

Katka Veselovská

Other topics of interest:

- construction grammar
- tectogrammatical description of English
- multimodal corpora
- automated metaphora detection and classification
- teaching: Linguistic Applications (FF UK, FF UPOL)
- theses consultations & supervisions
- business applications of text mining

http://ufal.mff.cuni.cz/~veselovska/

http://ufal.mff.cuni.cz/~seance/

Jonáš Vidra (vidra@ufal..., www.jonys.cz)

Master student of linguistics

thesis	Segmentation of words into morphemes (... using data from DeriNet)
supervisor	Zdeněk Žabokrtský

Other projects and iterests

- Machine learning: prediction of derivations in DeriNet
- Web technologies: Search engine for DeriNet

Zdeněk Žabokrtský: TEACHING

- courses taught in 2017/2018:
  - MFF UK: **Technology for NLP** (with Rudolf Rosa)
  - MFF UK: **Language Data Resources** (with Martin Popel)
  - MFF UK: **Machine Learning Methods** (with Ondřej Bojar)
  - FEL ČVUT: **Introduction to Natural Language processing** (with Jan Hajič, Dan Zeman, Pavel Pecina, Ondřej Bojar and Jindřich Libovický)
- Mgr. students supervised in 2017/2018
  - Jonáš Vidra, Josef Válek
- PhD. students supervised in 2017/2018
  - Martin Popel, Michal Novák, Rudolf Rosa, Nikita Mediankin, Vojtěch Hudeček

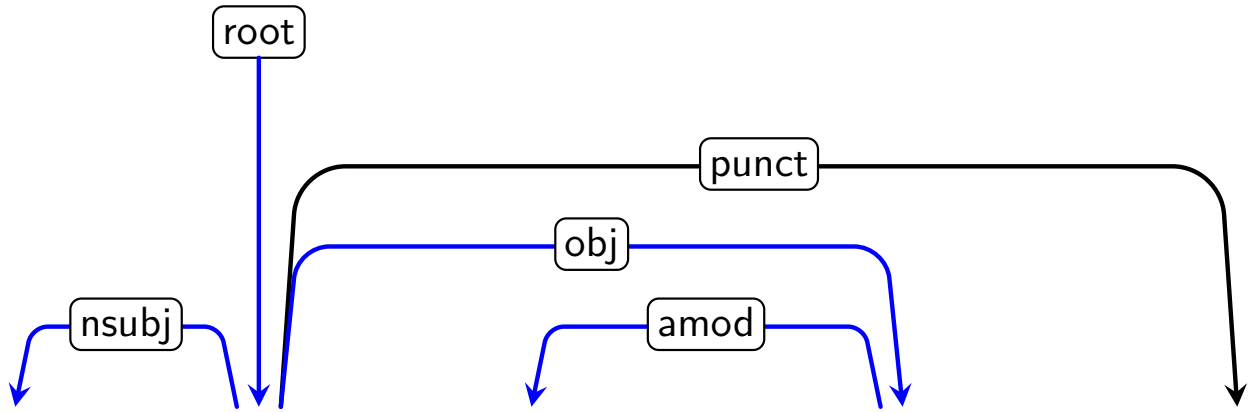
Zdeněk Žabokrtský: RESEARCH INTERESTS

- past
  - valency, treebanking, parsing, named entities, anaphora resolution ...
- current
  - ML applied in NLP
  - derivational morphology
  - dependency trees accross languages
  - in general: my research interest $=\cup$ research interests of my students

Zdeněk Žabokrtský: OFFICE

- chair of the board for the UFAL's PhD study program 4I3 Mathematical linguistics
- academic projects:
  - LangTech – a Ministry of Education project aimed at modernizing UFAL's PhD study program (PI)
  - DigiLing – an Erasmus+ international project (holder of the CUNI MFF+FF's part)
- recent/current research for non-academic partners:
  - Police of the Czech Republic
  - ACREA CZ
- academic service:
  - an evaluator in the National Accreditation Office
  - an evaluator in the Czech Technological Agency
  - all kinds of reviewing ...

Dan Zeman



PRON	VERB	ADJ	NOUN	PUNCT
PronType=Prs Case=Nom Number=Sing Person=1	VerbForm=Fin Mood=Ind Tense=Pres Number=Sing Person=1		Number=Plur	

Dan Zeman

- I am in the core group that coordinates the UD project

Dan Zeman

- I am in the core group that coordinates the UD project
- I have designed most of the morphological features in UD (⇐ Interset)

Dan Zeman

- I am in the core group that coordinates the UD project
- I have designed most of the morphological features in UD
  ($\Leftarrow$ Interset)
- I am responsible for final checks and releases of UD data in Lindat

Dan Zeman

- I am in the core group that coordinates the UD project
- I have designed most of the morphological features in UD (⇐ Interset)
- I am responsible for final checks and releases of UD data in Lindat
- I have converted the Czech data from Prague style to UD 
  - Prague Dependency Treebank
  - Czech Academic Corpus
  - Czech Legal Text Treebank
  - working on Czech Fiction Treebank (FicTree)
  - Also converted Polish  , Slovak  , Arabic  , Tamil  , Spanish  , Catalan  , Latin 
  - Significantly improved German  , Spanish  , Croatian 
  - Manually annotated Czech  and Upper Sorbian 

Dan Zeman

- I am in the core group that coordinates the UD project
- I have designed most of the morphological features in UD (⇐ Interset)
- I am responsible for final checks and releases of UD data in Lindat
- I have converted the Czech data from Prague style to UD
  - Prague Dependency Treebank
  - Czech Academic Corpus
  - Czech Legal Text Treebank
  - working on Czech Fiction Treebank (FicTree)
  - Also converted Polish, Slovak, Arabic, Tamil, Spanish, Catalan, Latin
  - Significantly improved German, Spanish, Croatian
  - Manually annotated Czech and Upper Sorbian
- Trying to coordinate efforts to improve consistency of UD data
- (Co-)organized the CoNLL 2017 shared task in parsing UD