Short presentations

ÚFAL seminar
Příchovice
19. – 22. 9. 2013

Last update: 18. 9. 2013 23:30

Contents

Paraphrasing Czech Sentences for MT Evaluation

Aim of the thesis: Improving the quality of n-gram based MT evaluation metrics by finding a paraphrase of the reference sentence that is closer in wording to the machine output than the original reference.

Machine translation:	**Banky** *zkoušejí platbu pomocí mobilního telefonu*
Reference translation:	**Banky** *testují placení mobilem*

Table: Example from WMT12

- **Done:** Simple greedy algorithm for lexical paraphrase substitution based on WordNet and Meteor tables.
- **Future work:**
  - grammar check
  - contextual validation
  - multi-word paraphrases
  - paraphrasing forms
  - word order changes
  - passive × active constructions

Eduard Bejček

Vallink

- linking valency frames from VALLEX to PDT-VALLEX

MWE identification [Pavel S., Pavel P.]

- searching for syntactic structures of known MWEs in a parsed text

ElixirFM – Arabic valency lexicon

- typesetting a book [Viktor Bielický]

Help and Support

- **(Xe)(La)TeX**
- **Russian and Polish valency lexicon** [dr. Skwarska]
- **annotation of light verb constructions** [Vendula Kettnerová]
- **searching in PDT** [prof. Panevová]
- **maybe some help with PDT release 3.0**

Silvie Cinková

- Semantic Pattern Recognition
  – with Martin Holub and Ema Krejčová

- Automatic Illustration
- Text in the Wild

Center for Large Scale Multi-modal Data Interpretation

Semantic Pattern Recognition

Martin Holub, Silvie Cinková, Ema Krejčová

- Formal model of selectional preferences of English verbs and nouns
  - Computing distributional similarity of nouns as verb complementizers
  - Relabeling of automatic a-layer to optimize dependency-based collocate extraction

Petra Galuščáková

- Information retrieval in audio-visual corpora
  - Topical segmentation of audio-visual recordings
  - Applying IR on the segments, using transcripts
  - GAUK
- **MediaEval** shared tasks
  - Benchmarking for processing of multimedia data
  - 2012: **Search and Hyperlinking** Task (semi-professional news videos)
  - 2013: **Similar Segments in Social Speech** Task (recorded dialogues) and **Search and Hyperlinking** Task (BBC programmes)

Similar Segments in Social Speech

- Query given as a segment in the recording
- Find more similar segments
- Given human and ASR transcripts, prosodic features, manually marked segments (and corresponding similar segments)
- ML-based searching for segment beginning and segment end (employing cue words and tags n-grams, silence, case, lexical chains, ...)

- Machine Translation
  - **WMT 2013**: Statistical post-editing of the TectoMT output and using synthetic parallel data from TectoMT to train a statistical MT system (with Martin Popel and Ondřej Bojar)
- Center for Large-Scale Multi-modal Data Interpretation
  - Collecting and cleaning news articles with corresponding images from Profimedia dataset
  - Downloaded 50k articles and 130k images from 23 sources

EU projects at UFAL

Jan Hajič

Institute of Formal and Applied Linguistics

Computer Science School

Charles University in Prague

Czech Republic

UFAL seminar Prichovice

Overview

- Running
  - Khresmoi (2010-2014)
  - Eudat (2011-2014)
  - PARSEME (2013-2017)
  - LINDAT/Clarin (2010-2015)
- New / upcoming
  - QTLeap (2014-2017)
  - PARSEME CZ (2014-2017)
  - JHU Workshop 2014 (2014)
- Submitted
  - Center of Excellence (GACR, 2014-2018)
  - Center of Competence (TACR, 2014-2019)
- Horizon 2020

Overview

- Running
  - Khresmoi (2010-2014)
  - Eudat (2011-2014)
  - PARSEME (2013-2017)
  - LINDAT/Clarin (2010-2015)
- New / upcoming
  - QTLeap (2014-2017)
  - PARSEME CZ (2014-2017)
  - JHU Workshop 2014 (2014)
- Submitted
  - Center of Excellence (GACR, 2014-2018)
  - Center of Competence (TACR, 2014-2019)
- Horizon 2020

Khresmoi (2010-2014)

http://www.khresmoi.eu

- Machine translation for CLIR
  - Cross-language information retrieval
    - Query: Czech, French, German → English, documents in Eng
  - Domain adaptation (medical domain)
  - “Genre” adaptation (short quearies)
  - Paper accepted: AIIM journal (P. Pecina et al., with DCU)
- User evaluation
  - Subjective full system task-based evaluation, incl. UI
    - First experience at UFAL (Z. Uresova)
  - First round: May/June 2013/14; report available
- Data collection
  - MT, Czech medical websites for full system

Eudat (2010-2014)

- Analogy to Clarin preparatory phase
  - Prepare European research infrastructure for sceintific data (of any kind, incl. LR)
  - Tübingen and us – language community representatives
  - http://www.eudat.eu
- Goals
  - (low-level) standards for data storage, replication, identification, IPRs, authentication for storage and access, search, web services
  - Our goal: compatibility with Clarin
    - Future: Czech large infrastructure

PARSEME (2013-2017)

- New EU networking project (started March 2013)
- Single topic: Multi-word Expressions, 4 areas:
  - MWE representation
  - MWEs in parsing (improving parsing/MWE)
  - MWEs and hybrid parsing
  - MWEs and annotation/treebanks
- UFAL work:
  - MWE in PDT 2.5, MWEs & valency
  - E. Bejcek, P. Stranak

LINDAT/Clarin (2010-2015)

- Three tasks
  - Building a repository (J. Misutka+4)
    - dSpace adaptations, ready for certification
  - Czech data preparation (→ repository)
    - Annotation (new PDT/discourse/MWEs/...)
  - Admin/coordination
    - Clarin EU conference in Prague, Oct. 13)
- New focus: web services and applications
  - PML-TQ, treex on the web, demos, ...
  - Applications for humanities research
- Going “operational” 1/1/2014

Overview

- Running
  - Khresmoi (2010-2014)
  - Eudat (2011-2014)
  - PARSEME (2013-2017)
  - LINDAT/Clarin (2010-2015)
- New / upcoming
  - QTLeap (2014-2017)
  - PARSEME CZ (2014-2017)
  - JHU Workshop 2014 (2014)
- Submitted
  - Center of Excellence (GACR, 2014-2018)
  - Center of Competence (TACR, 2014-2019)
- Horizon 2020

QTLeap (2014-2017)

- The last MT project funded by the 7th FP EU
  - 8 partners
    - coordination: Univ. of Lisbon, António Branco
- Goal: “high-quality” translation
  - By incorporation of “linguistic” features
    - Esp. semantics
  - 3 years, 3 prototypes (“pilots”)
    - baseline, advanced, “semantic”
  - Spanish, Portuguese, German, Czech, Dutch, Bulgarian, Basque

PARSEME CZ (2014-2017)

- Multiword entities
  - Czech, possibly other languages
  - Annotation, extraction, use in NLP tasks
    - MWEs incl. verbal ones
- Separate from PARSEME (EU)
  - Czech funding only
  - E. Bejček, P. Straňák
- (Submitted only...!)

JHU Workshop 2014

- Prague 2014 (7.7.-1.8.2014)
  - Due to „colored“ money by the NSF…
- 2 teams
  - Speech (JHU, VUT Brno, others)
  - Translation (semantic/AMR)
    - UCB, Brandeis, Rochester (Dan Gildea), UFAL
- First week: “PRELIM” workshop/school
  - Lectures on “pie-in-the-sky” topics
    - Neurolinguistics, cogsci, methods in biology, advanced unsupervised techniques, ...
  - Organized by Jason Eisner (JHU)

Overview

- Running
  - Khresmoi (2010-2014)
  - Eudat (2011-2014)
  - PARSEME (2013-2017)
  - LINDAT/Clarin (2010-2015)
- New / upcoming
  - QTLeap (2014-2017)
  - PARSEME CZ (2014-2017)
  - JHU Workshop 2014 (2014)
- Submitted
  - Center of Excellence (GACR, 2014-2018)
  - Center of Competence (TACR, 2014-2019)
- Horizon 2020

Submitted projects

(some small chance...)

- Center of Excellence (GACR, 2014-2018)
  - Focus on: speech and spoken language
    - Spoken language studies (“linguistics”)
    - Search (~ LINDAT/Clarin) in spoken/audio data
    - ASR and TTS (Pilsen)
    - Information Extraction incl. Sentiment Analysis
    - Dialog Management
  - Pilsen, UJC, UFAL
- Center of Competence (TACR, 2014-2019)
  - Applied research on
    - Speech and text technologies for applications
    - Information mining for business support
  - VUT Brno, ZCU Plzen, UFAL, Lingea, Phonexia, iCord

Overview

- Running
  - Khresmoi (2010-2014)
  - Eudat (2011-2014)
  - PARSEME (2013-2017)
  - LINDAT/Clarin (2010-2015)
- New / upcoming
  - QTLeap (2014-2017)
  - PARSEME CZ (2014-2017)
  - JHU Workshop 2014 (2014)
- Submitted
  - Center of Excellence (GACR, 2014-2018)
  - Center of Competence (TACR, 2014-2019)
- Horizon 2020

Horizon 2020

- New EU program for research 2014-2020
- Three general areas again (plus infrastructures):
  - Fundamental research (ERC)
  - Future and emerging technologies (FET)
  - Standard research projects
    - Little technological projects like MT
      - 1st call 2014:
        - 15 MEUR only for “cracking the language barrier” aka MT
        - Some robotics topics with speech/NLP?
        - eHealth?
    - Applications: “Societal Challenges”
      - Health, Food, Energy, Transport, Climate, IIS (Identity/Heritage), Security
- Task for all: look for partners!

Thank you!

Jarka Hlaváčová

- Aktualizace morfologického slovníku češtiny
  - opravy, nová slova, kontroly
  - nejčastější náměty a požadavky na doplnění z ÚFAL, ÚTKL, ÚČNK
- „Morfologický servis“
  - interní
  - pro ÚČNK, ÚTKL

- příprava dat pro projekt Khresmoi (biomedicínská doména)

- výuka: Perl pro neinformatiky – společně pro FF UK

Filip Jurčíček

- Main activities in the last year
  - Teaching:
    - STATISTICAL DIALOGUE SYSTEMS
      - Lectures - 2 hours per week
      - Students were building their own components for an real SDS
    - BAYESIAN INFERENCE
      - Lectures - 2 hours per week
      - This year – invited lecturers from Cambridge, UK
    - MASTER THESES
  - Projects:
    - VYSTADIAL → Research into spoken dialogue systems
    - FRVS Grant → BAYESIAN INFERENCE course

Master theses

- DEFENDED - David Marek
  - Bayesian inference for belief tracking is SDS

    - participated in the 2012 Dialogue State Tracking Challenge

- TO BE FINISHED – Ondrej Oplatek
  - KALDI Online ASR Decoder

VYSTADIAL

- PI: Filip Jurčíček
  - Development of statistical methods for spoken dialogue systems
  - 1.4.2012 – 31.12.2016

- Funding for 4 PhD students
  - one place still available

- Collaborators:
  - Ondrej Dusek, Matej Korvas, Lukas Zilka, Ondrej Platek, ~~David Marek~~, Filip Sedivy

Main Activities

- 2012 Dialog State Tracking Challenge
  - Duration: Jan – March 2013
  - A publication at SigDial 2013
  - Designed and evaluated two dialogue state trackers

- Deployment of Public Transport Info system prototype
  - 800 899 998 phone number
  - You can ask about public transport connections in Prague
  - This task should be our test bed in Czech language
  - We are collecting data to train first version of our statistical components

•Kettnerová Václava

- **Grants**
- Delving deeper: Lexicographic description of syntax and semantic properties of Czech verbs (M. Lopatková)
- Computational Linguistics: Explicit description of language and annotated data focused on Czech (J. Panevová)

- **Research interest**
- valency of verbs, alternations – changes in valency structure, lexical-semantic representation, semantic classification of verbs
- lexicographic representation of changes in valency structure
- now: **light verbs:** btw. nouns and verbs; derivation of surface structure of light verb constructions, their lexicographic representation and detection
- in the future: derivational relations of verbs with respect to changes in valency

Natalia Klyueva

- Ph.D. Student, 7th year
- Linguistic aspects of Machine Translation between Czech and Russian
- Project: GAUK 639012 (2012-2013) **Machine Translation between related languages**, with Karel Bílek and Vladislav Kuboň:
  - Main question: which approach - RBMT or SMT - is more suitable for the related languages
  - Analyzing error types: whether errors can be justified by the system settings/architecture/data or by the fact of the discrepancies between the languages
- **Done so far,** in collaboration with other ÚFAL colleagues: Česílko (Petr Homola), TectoMT (Martin Popel a Zdeněk Žabokrtský), Moses
- RBMT gives better sentence structure, SMT performs better in terms of semantics
- Error assignment:

```
*SRC Odpověď je jasná : americký systém vede na celé čáře .
MOS ответ очевиден : американская система ведет idiom::{в результате} .
GOO Ответ очевиден: американская система disam::приводит idiom::{к линии}.
PCT Ответ svagr_gen::понятна : американский комплекс disam::ведает idiom::{на целой полосе}
CSL Ответ be::они agr_gen::ясная : pos::сша система ms::приводит idiom::{на весь разделение}
TMT unk::Одповед  agr::ясная : американская система conj::вести idiom::{на всей линии} .
```

Natalia Klyueva

- Improve Moses, brute force approach – add respective data based on the error analysis
- Start Cze-Ru pair in Apertium (Česílko-like RBMT)
- Low-priority tasks, just ideas:
  - RuVallex – valency discrepancies between Czech and Russian
  - Make a parallel between bilingualism/2LL and Machine Translation
  - Error correction interface, not only for MT

Veronika Kolářová

- Mgr.: FF UK (Czech & Serbian / Croatian; 1998)
- Ph.D.: UFAL MFF UK (Valency of nouns; 2006)

Participation in two GAČR projects:

- Systematic, economical and corpus-based description of valency properties of Czech deverbal nouns (theory and practice) P406/12/P190
  - post-doctoral project, 2012-2014
  - principal investigator
- Computational Linguistics: Explicit description of language and annotated data focused on Czech P406/10/0875
  - standard project, headed by prof. Panevová; 2010-2013
  - team member

- Main topics of interest:
  - valency of Czech deverbal nouns
  - support verb constructions

Systematic, economical and corpus-based description of ***valency*** *properties of* ***Czech deverbal nouns*** *(theory and practice)*

GA ČR P406/12/P190

- Post-doctoral project
- Principal investigator: Veronika Kolářová
- 2012-2014
- Total financial support (3 years): 1 499 000 CZK

The goals of the project

Theory:

- To complete the description of adnominal counterparts of adverbal objects expressed by **prepositionless cases**
  - especially genitive and instrumental
    - *jeho dotek puku* ‘he-POSS.SG touch-NOM.SG puck-GEN.SG’
    - *nákaza chřipkou* ‘infection-NOM.SG flu-INS.SG’
- To specify unique valency properties of some semantically compact groups of nouns
  - nouns of communication, nouns of exchange, nouns denoting mental state or dispositions

Practice:

- To incorporate knowledge about nominal valency into
  - the treatment of dictionary entries (PDT-VALLEX)
  - the guidelines for annotation of PDT (not annotation as such)

Current work

- Agents expressed by prepositionless instrumental [A_1(Ins)] modifying Czech nouns derived from **intransitive verbs**
  - – modification by A_1(Ins) occurs in CNC subcorpora with
  - – (i) nouns derived from verbs that can be passivized
    - *vyhrožování* *rozhodčím*.ADDR *trenérem*.ACT (SYN2006PUB)

      threatening referee-DAT.PL coach-INS.SG

      ‘threatening to the referees by the coach’
  - – (ii) nouns the source verbs of which cannot be changed to passive
    - especially nouns derived from reflexive verbs
    - *zmocnění se* *televize*.PAT *teroristy*.ACT (SYN2009PUB)

      seizure REFL television-GEN.SG terrorist-INS.PL

      ‘seizure of the television by terrorists’

Current work

- Modification by A_1(Ins) is possible even when the second complementation A_2 is omitted on the surface
  - *Po domluvě strážníky*.ACT *děti z místa odešly.* (SYN2009PUB)
    ‘After caution by police officers children leaved the place.’
  - *jakékoli napomáhání sestřičkou*.ACT *je … vyloučeno.* (SYN2009PUB)
    ‘any helping by the nurse is … excluded’
  - *klasické vloupání neznámým pachatelem*.ACT. (SYN2000)
    ‘classic break-in by an unknown perpetrator’

- Accepted for the SLOVKO conference (Bratislava, November 2013)

SLU in Alex

Matěj Korvas

Monday 9th September, 2013

Outline

The Alex dialogue system

SLU before Alex

SLU in Alex

References

Alex – Overview

Project Alex is led by Filip Jurčíček. The goal is to implement a generic statistical spoken dialogue system.

Alex is based on Filip's previous work in Cambridge.

Domains:
- TownInfo – tourist information about bars, hotels, etc.
- CamInfoRest – ditto, but only for dining venues
- Alex On The Bus – finding public transport connection in Prague

Alex – Overview

Project Alex is led by Filip Jurčíček. The goal is to implement a generic statistical spoken dialogue system.

Alex is based on Filip's previous work in Cambridge.

Domains:

- TownInfo – tourist information about bars, hotels, etc.
- CamInfoRest – ditto, but only for dining venues
- Alex On The Bus – finding public transport connection in Prague

Alex – Overview

Project Alex is led by Filip Jurčíček. The goal is to implement a generic statistical spoken dialogue system.

Alex is based on Filip's previous work in Cambridge.

Domains:
- TownInfo – tourist information about bars, hotels, etc.
- CamInfoRest – ditto, but only for dining venues
- Alex On The Bus – finding public transport connection in Prague

Alex – Overview

Project Alex is led by Filip Jurčíček. The goal is to implement a generic statistical spoken dialogue system.

Alex is based on Filip's previous work in Cambridge.

Domains:

- TownInfo – tourist information about bars, hotels, etc.
- CamInfoRest – ditto, but only for dining venues
- Alex On The Bus – finding public transport connection in Prague

Alex – Overview

Project Alex is led by Filip Jurčíček. The goal is to implement a generic statistical spoken dialogue system.

Alex is based on Filip's previous work in Cambridge.

Domains:

- TownInfo – tourist information about bars, hotels, etc.
- CamInfoRest – ditto, but only for dining venues
- Alex On The Bus – finding public transport connection in Prague

Alex – Overview

Project Alex is led by Filip Jurčíček. The goal is to implement a generic statistical spoken dialogue system.

Alex is based on Filip's previous work in Cambridge.

Domains:

- TownInfo – tourist information about bars, hotels, etc.
- CamInfoRest – ditto, but only for dining venues
- Alex On The Bus – finding public transport connection in Prague

Outline

The Alex dialogue system

SLU before Alex

SLU in Alex

References

DA Representation

DA (dialogue act) is represented as a conjunction of DAIs.

dai&dai&dai

DAI (dialogue act item) represents a unit of interaction between the user and the system.

da_type(slot_name = slot_value)

Examples of DAIs:

- `confirm(area="girton")`
- `inform(="pub")`
- `request(address)`
- `bye()`

Semantic Tuple Classifier

- Used in Cambridge systems [Mairesse et al., 2009].
- DAI is a ⟨DA_type, slot_name, slot_value⟩ tuple.
- One SVM for smaller sub-tuples, complete DAIs re-built from these.
- Slot value abstraction: `Girton` ↦ `AREA-0`.

Outline

The Alex dialogue system

SLU before Alex

SLU in Alex

References

Filip's SLU implementation

Filip's original implementation

Let's do it simply: one MaxEnt classifier for each *complete DAI*.

Results: DAI F-score ∼95% in the given domain.

Room for improvement

- Process more informative output of ASR (*n-best lists, confusion networks*).
  (In Cambridge, they were faster to implement this: [Henderson et al., 2012].)
- Improve slot value abstraction – avoid category labels `AREA-0`, `AREA-1`, `AREA-2`, . . .

Filip's SLU implementation

Filip's original implementation

Let's do it simply: one MaxEnt classifier for each *complete DAI*.
Results: DAI F-score ∼95% in the given domain.

Room for improvement

- Process more informative output of ASR (*n-best lists, confusion networks*).
  (In Cambridge, they were faster to implement this: [Henderson et al., 2012].)
- Improve slot value abstraction – avoid category labels `AREA-0`, `AREA-1`, `AREA-2`, . . .

My Contribution

probabilistic representation

I implemented the preprocessing and n-gram extraction from n-best lists and confnets.

Tricky points:

- Confnet with category symbols (e.g., `FOOD-0`) substituted for multiword phrases ceases to be a confnet.

My Contribution

probabilistic representation

I implemented the preprocessing and n-gram extraction from n-best lists and confnets.

Tricky points:

- Confnet with category symbols (e.g., `FOOD-0`) substituted for multiword phrases ceases to be a confnet.
  → implemented a relaxed version of confnet

My Contribution

probabilistic representation

I implemented the preprocessing and n-gram extraction from n-best lists and confnets.

Tricky points:

- Confnet with category symbols (e.g., `FOOD-0`) substituted for multiword phrases ceases to be a confnet.
  → implemented a relaxed version of confnet
- Confnets output by ASR typically contain many empty words.

My Contribution

probabilistic representation

I implemented the preprocessing and n-gram extraction from n-best lists and confnets.

Tricky points:

- Confnet with category symbols (e.g., `FOOD-0`) substituted for multiword phrases ceases to be a confnet.
  → implemented a relaxed version of confnet
- Confnets output by ASR typically contain many empty words.
  → n-grams not always occupy the same span of the confnet

My Contribution

category labels without numbering

Category labels were introduced to defeat sparsity. **However**, they are numbered! This reintroduces the slot value sparsity.

My Contribution

category labels without numbering

Category labels were introduced to defeat sparsity. **However**, they are numbered! This reintroduces the slot value sparsity.

→ Need to turn around the entire learning process – for each training example, *instantiate* its category labels only when we know what category we train for.
. . . Work with *class-dependent features*.

My Contribution

category labels without numbering

Category labels were introduced to defeat sparsity. **However**, they are numbered! This reintroduces the slot value sparsity.

In other words:

1. category label substitution & feature extraction
2. training classifiers

becomes

1. category label substitution (without numbering)
2. for each DAI
   - extract features for this category
   - train the classifier

Outlook

- introduce other features than just word n-grams
- relax the category assignment from exact match to phonetic, or perhaps semantic similarity
- go incremental

Outlook

- introduce other features than just word n-grams
- relax the category assignment from exact match to phonetic, or perhaps semantic similarity
- go incremental

Outlook

- introduce other features than just word n-grams
- relax the category assignment from exact match to phonetic, or perhaps semantic similarity
- go incremental

- Henderson, M., Gasic, M., Thomson, B., Tsiakoulis, P., Yu, K., & Young, S. (2012).
  Discriminative spoken language understanding using word confusion networks.
  In *SLT* (pp. 176–181).

- Mairesse, F., Gasic, M., Jurcicek, F., Keizer, S., Thomson, B., Yu, K., & Young, S. (2009).
  Spoken language understanding from unaligned data using discriminative classification models.
  In *IEEE proceedings* (pp. 4749–4752).: IEEE.

Vláďa Kuboň

- Project
  - LCT – project has been newly approved by EU in 2012, this year there was a transition period during which it was necessary to reduce the debt of 300 000 EUR caused by the first coordinator Valia Kordoni
- Research
  - Syntactic analysis
    - Segmentation of complex sentences, formal properties of free word order, analysis by reduction
  - MT between related languages
  - Program comittee of several workshops (EAMT, IIS, BSNLP, FLAIRS etc.)
- Teaching
  - 2 lectures - Introduction to CL and NLP Applications;
2 seminars for UFAL and
1 seminar of Automata and Grammars for CS students
  - UFAL secretary for teaching
  - Coordination of an Erasmus exchange with Saarbruecken, Koper and Tuebingen
  - Supervising 1 thesis (defense in January),
3 PhD. students

Jindřich Libovický

- finished master’s studies at ÚFAL in the summer
  - thesis: Statistical NLP Method in Music Notation Analysis
  - talk on this topic on Monday seminar in November
- now starting doing the PhD (supervisor Pavel Pecina)
  - working on GAČR project: Center for large-scale multi-modal data interpretation
  - language modeling for text recognition in the real-world images

Markéta Lopatková - Projects

Research interests / research projects:

- Valency lexicon of Czech verbs – VALLEX
  esp. with Václava Kettnerová (past - Zdeněk Žabokrtský)
  diatheses and alternations
  enriching the lexicon with semantic information

  GAČR (2012-2015): *Delving Deeper: Lexicographic Description of Syntactic and Semantic Properties of Czech Verbs*
  (1.2 full contract)

- 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)

  GAČR: *NoSCoM: Non-Standard Computational Models and Their Applications in Complexity, Linguistics, and Learning*, 2010-2014
  (bonuses)

Delving Deeper: Lexicographic Description of Syntactic and Semantic Properties of Czech Verbs

- changes in valency structure of verbs, their representation in a lexicon
  - theoretical research; design of a formal model for lexicographic description
  - grammaticalized alternations: diatheses and reciprocity
  - lexicalized alternations: theoretical and practical aspects
  - comparative aspects of diatheses
  - application in an electronic language resource
- mapping lexical resources:
  - enhancing Czech valency lexicon with semantic classes and semantic roles; based on FrameNet
  - strengthening lexical resources with corpus evidence (VALEVAL)

Delving Deeper: Lexicographic Description of Syntactic and Semantic Properties of Czech Verbs

- GA P406/12/0557, duration 2012-2015
- budget: 7.137 mil. CZK
- partners:
  - ÚFAL:

    Markéta Lopatková, Vendula Kettnerová, Eda Bejček, Anša Vernerová (1.2 contract)
- Institute of Slavonic Languages, Academy of Science of the Czech Republic:

  Karolína Skwarska (0.7 full contract)

NoSCoM: Non-Standard Computational Models and Their Applications in Complexity, Linguistics, and Learning

- GA P202/10/1333, duration 2010-2014
- Institute of Computer Science, Academy of Science of the Czech Republic:
  Jiří Šíma, Jiří Wiedermann, Petr Savický, Stanislav Žák, Robert Kessl
- MFF UK:
  Martin Plátek, Markéta Lopatková, Fero Mráz, Iveta Mrázová, Peter Černo
- topics:
  1. Unconventional Computational Models
  2. Neural Networks
  3. Specialized Unbounded Automata and Grammars
     - modeling of stratificational dependency-based syntax
     - based on the analysis by reduction and restarting automata
     - recently, focus on free word-order:
       (non-)projectivity of a sentence and a number of word order shifts
     - RA and PDT
     - model of a lexicon
  4. Branching Programs

Central Funding

- PROVOZ (teaching money)
  - ca 1.6 mil. CZK salaries (3.5 full contracts)
  - 200 th. CZK other costs
- PRVOUK (research money)
  - ca 3.2 mil. salaries (6.1 full contracts)
  - 650 th. CZK other costs
- Specific Research (?)
  - 145 th. CZK other costs

Markéta Lopatková – Teaching (1)

"Teaching projects":

- Accreditation

- EM LCT (Language and Communication Technologies)
  together with Vladislav Kuboň
  2 students for 2013-14 (+ 1 scholarship)
  - new phase: 2013-2018?
    selected for funding (at least 5 students for 2014/15)

- CLARA (Common Language Resources and their Applications
  Marie Curie Action, 2009-13

- involved in a preparation of BSc. "General Computer Science" in English (from 2013/14)

Markéta Lopatková – Teaching (2)

Courses:

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

Supervising:

- 4 PhD students, 1 Master students

Others:

- Grant Agency of Charles University
  committee for computer science (oborová rada)
- Czech Science Foundation / GAČR
  panel P406 *Linguistics and Literature*
- editorial board: *Slovo a slovesnost*, *Korpus – Gramatika – Axiologie*
- coordinator of Erasmus exchange: Bolzano, Malta, Utrecht, Groningen

David Mareček

Teaching:

Selected Problems in Machine Learning (with ZŽ)

- Bayesian inference
- Gibbs sampling

Research:

Unsupervised dependency parsing

- supervised POS tags, unsupervised POS tags

HamleDT

- parsing with different encoding of coordination structures
- separated parsing of coordinations

Khresmoi

- splitting German compounds for better MT and IR

Marie Mikulová

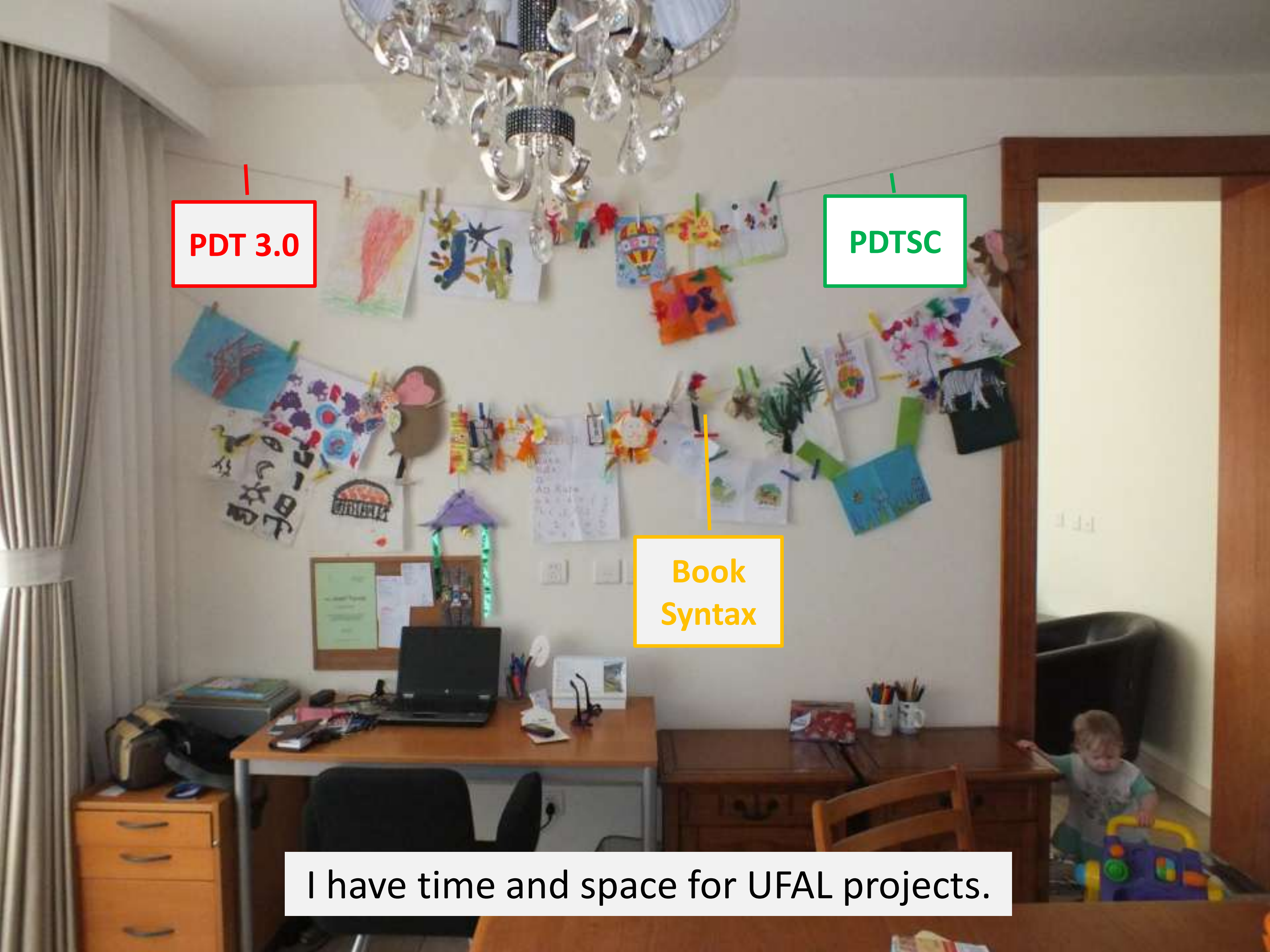
Many greetings from Beijing!

Long-term foreign bussiness trip, Beijing, China, till March 2015

My husband works at Volkswagen Group China.

Ludmila has her nanny (ayi).

Růžena and Ladislav attend Dulwich College Beijing.

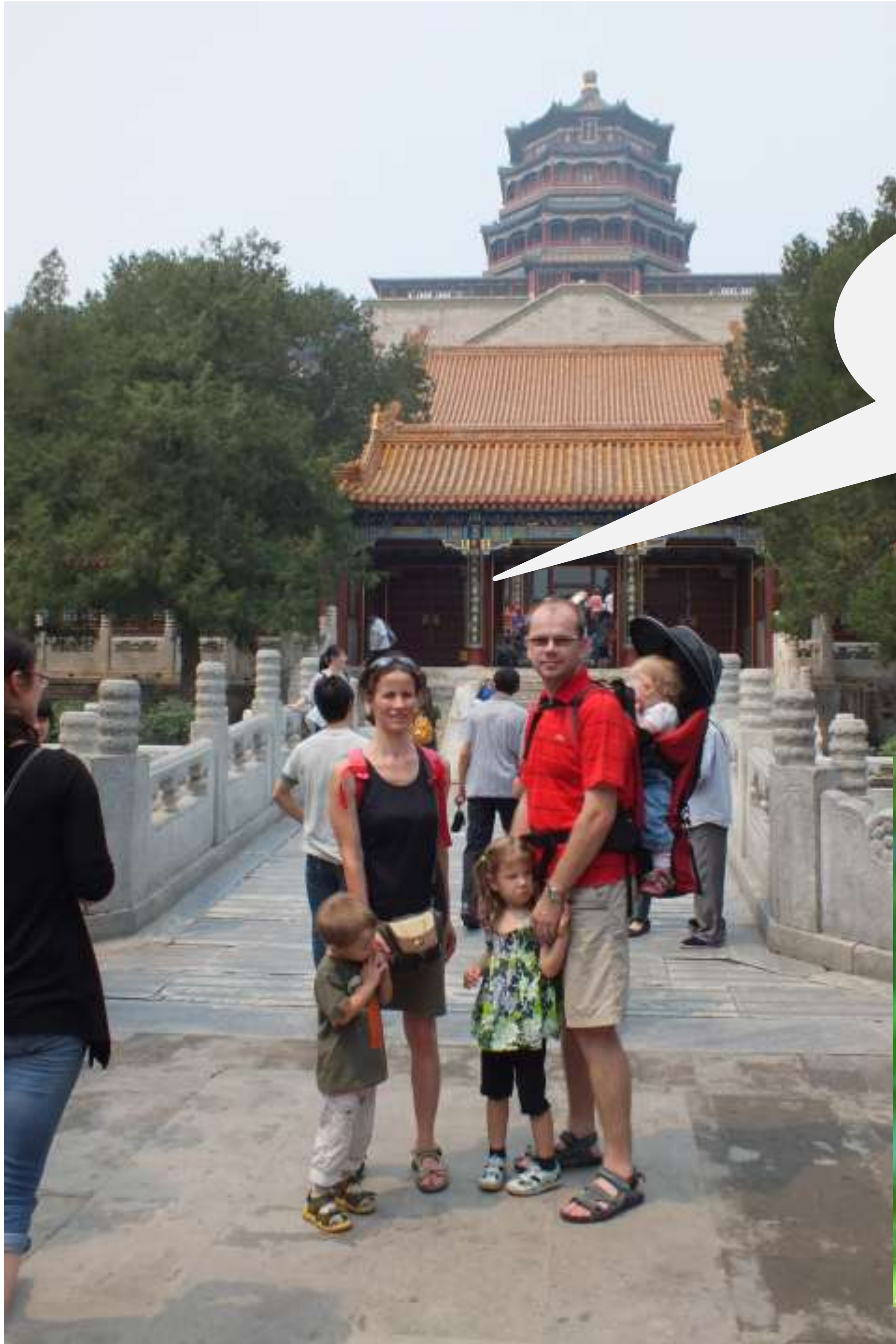


PDT 3.0

PDTSC

Book Syntax

I have time and space for UFAL projects.



See you
in Beijing!
☺

Jiří Mírovský / Discourse

Lucie Poláková, Pavlína Jínová, Magdaléna Rysová, Šárka Zikánová, Prof. Hajičová, and others:

Published in November 2012 as **Prague Discourse Treebank – PDiT** (PrDiT suggested but rejected)

To be published in **PDT 3.0** (updated version), already in svn PDT 2.x

Annotation of AltLex's (small adjustments to the tool)

Jiří Mírovský / Anaphora

Anja Nedoluzhko and others:

- **P**ublished in December 2011 (over PDT 2.0)
- **A**n updated version published in November 2012 as part of **PDiT** (over PDT 2.5)
- **T**o be published in **PDT 3.0** (+ 1st and 2nd person), not yet in svn PDT 2.x

Jiří Mírovský / tfa

Kateřina Rysová and others:

Annotation of contextual boundness (**tfa**) in **PCEDT** (5 thousand sentences on each side)

Czech part almost finished (5 th. sentences)

English part just started

Preannotation in Czech (a set of rules – 1/3 of nodes)

Preannotation in English (a set of rules + transfer from Czech – 4/5 of nodes)

Jiří Mírovský / Other stuff

Kateřina Rysová:

Word order in Czech

Prof. Panevová, Magda Ševčíková:

Changes in some attributes for **PDT 3.0**

Markéta Lopatková, Vladislav Kuboň:

Analysis by reduction performed on analytical trees

Příchovice, September 2013

Jiří Mírovský / Pub. service

Purchase and management of software (but not SW from Microsoft)

Purchase and management of data published by **LDC**

Assistant management of **Amoeba** (database of employees at ÚFAL)

Anja Nedoluzhko - coreference

- Prague Discourse Treebank 1.0 (GAČR Šárky Zikánové Coreference, discourse relations and information structure in a contrastive perspective + GAČR prof. Panevové)
- analysis of coreference relations in the annotated corpus, interplay with discourse and tectogrammatics
  - overviews: Nedoluzhko-Mírovský-Novák (coreference&resolution), Poláková et al. for IJCNLP (discourse)
  - generic NP coreference (Nedoluzhko for ACL, LAW)
  - coreference/bridging and tectogrammatics (Nedoluzhko-Mírovský for DepLing)
  - analysis of inter-annotator agreement (Nedoluzhko-Mírovský for DepLing)
  - translation of *it* (Novák-Nedoluzhko-Zdeněk for ACL, DiscoMT and IJCNLP)
  - #PersPron coreference?

Anja Nedoluzhko – other topics

- PEDT: (almost) finished annotation of NP-coreference for English, some controls remaining
- morphology: productive prefixation (Jarka-Nedoluzhko for TSD), planned to extend more to linguistics
- ??? morphological annotation of the Upper Sorbian language corpus

Michal Novák

- GAUK 4226/2011: Utilization of coreference in machine translation
  - improving translation of “it” and reflexive pronouns into Czech
    - with Anja and Zdeněk
  - coreference to help translating other words than pronouns
    - with Liane Guillou
- Khresmoi
  - data filtering
- public service: communicating with Karolinum bookstore

Nino Peterek

Projects

- Pronunciation features of Czech language - dialect analysis (with Vendula Michlíková).
- Continously learning analyser of audio-visual recordings (with Ondřej Košarko).
- Sound recognizer of particular grasshopper species (with Jan Schwarz).
- Speech corpora processing.

Teaching

- Fundamentals of speech recognition and generation.
- Natural computing for learning and optimisation.
- Algorithms in speech recognition (an advanced course).

Lucie Poláková

Project:
Annotation of discourse structure on PDT (discourse connectives, their scopes + meanings, textual coreference, bridging anaphora)

- CD with **PDiT 1.0 released** in November 2012
(summarizing paper documenting the PDiT release accepted for ICJNLP 2013)
- Genre specification of PDT texts – finished
- Enhanced version of discourse phenomena annotations ready for PDT 3.0

- Grant support: GAČR (Zikánová, “Interplays” till 2015), LINDAT

Recent about the project:

- Cooperation with UPenn and prof. Aravind Joshi’s team
– new application for extension of the KONTAKT grant
- Annotation of alternative lexicalizations in PDiT (Majda Rysová)

Lucie Poláková

Recent about the project:

ADACA: Advances in Discourse Analysis and its Computational Aspects (CoLing 2012, Eva Hajičová)

COST: New huge project on multilingual corpora with linked connectives

DiscoMT workshop at ACL: how phenomena "beyond the sentence boundary" can influence SMT, shared task – B. Webber, A. Popescu-Belis)

PhD: Dissertation topic:

The concept of discourse-level description for the PDT

Administrative:

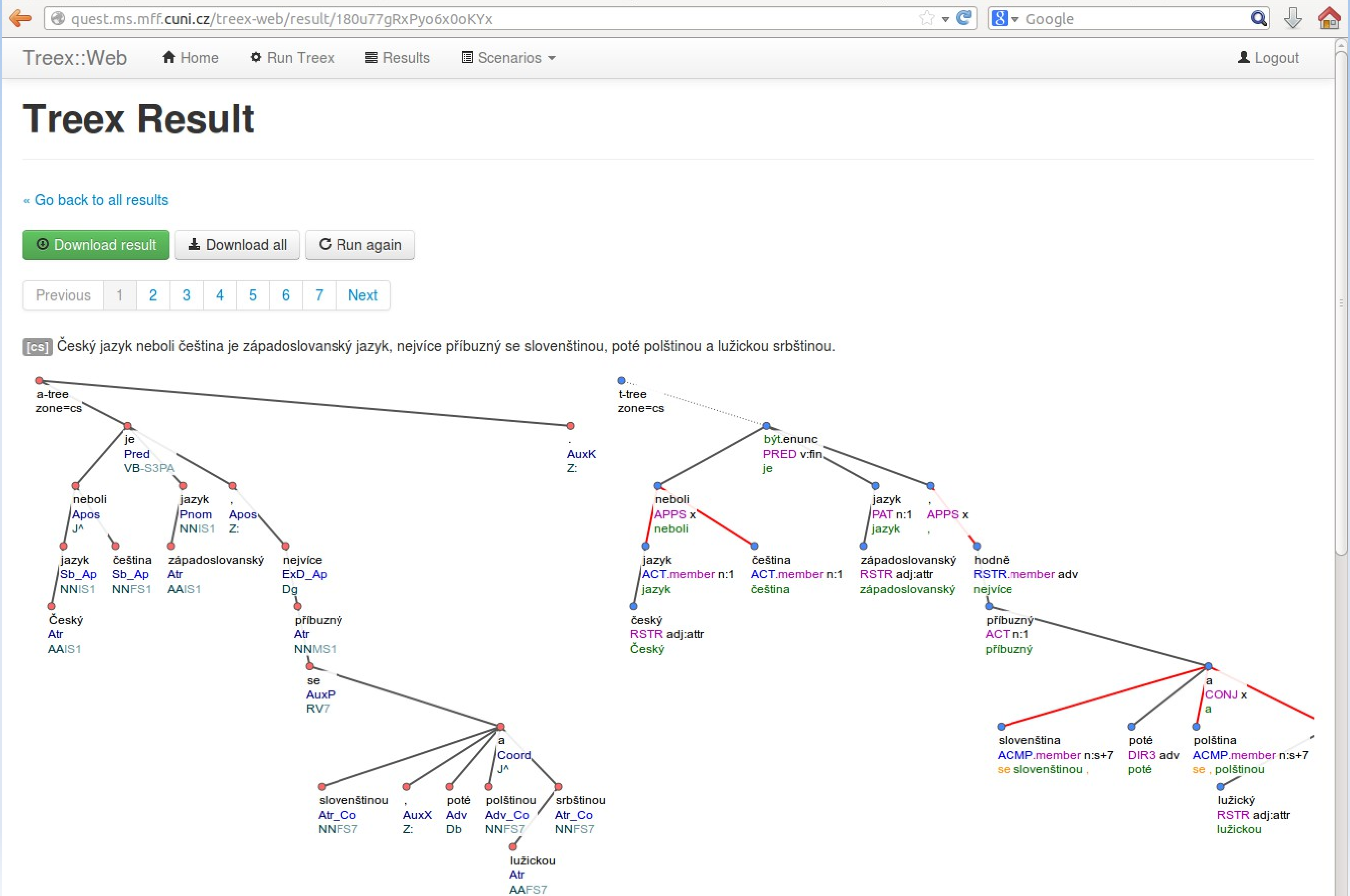
new UFAL webpages, 4th floor posters „nástěnkář“

Martin Popel

- **Treex** NLP framework (http://ufal.mff.cuni.cz/treex/)
  - Michal Sedlák: Treex::Web (Bc thesis, LINDAT demo)
- **TectoMT** machine translation (PhD thesis on transfer)
  - combinations with Moses (WMT13: Chimera, PhraseFix)
  - Ondřej Klejch: MTCompareEval (Bc thesis, to do: statmt.org)
- **HamleDT** 30+ treebanks (ACL13: coordinations)
- **PBML** (next deadline: January 8th 2014)
  - Matěj Korvas (LaTeX), Kateřina Stuparičová (admin.)
- **Technical reports** (2013 deadline: December 1st)
- **Teaching** (autumn: Modern Methods in CL “Reading group”)

Treex::Web

http://quest.ms.mff.cuni.cz/treex-web/

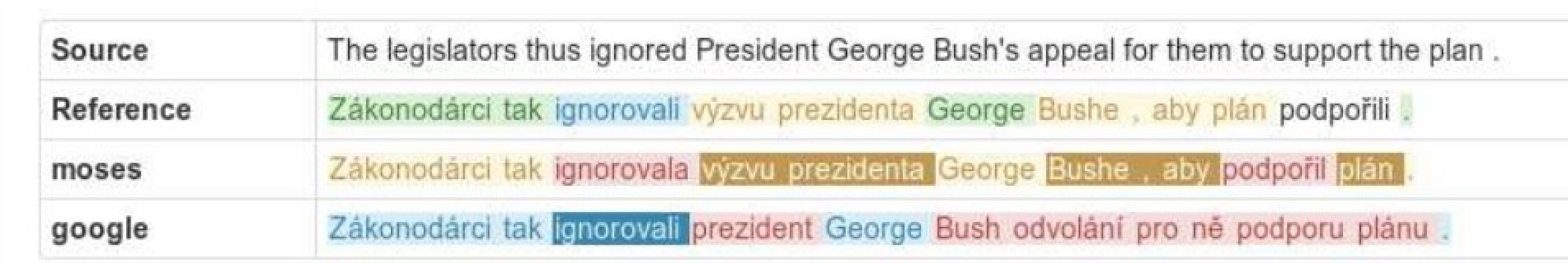


Treex Result

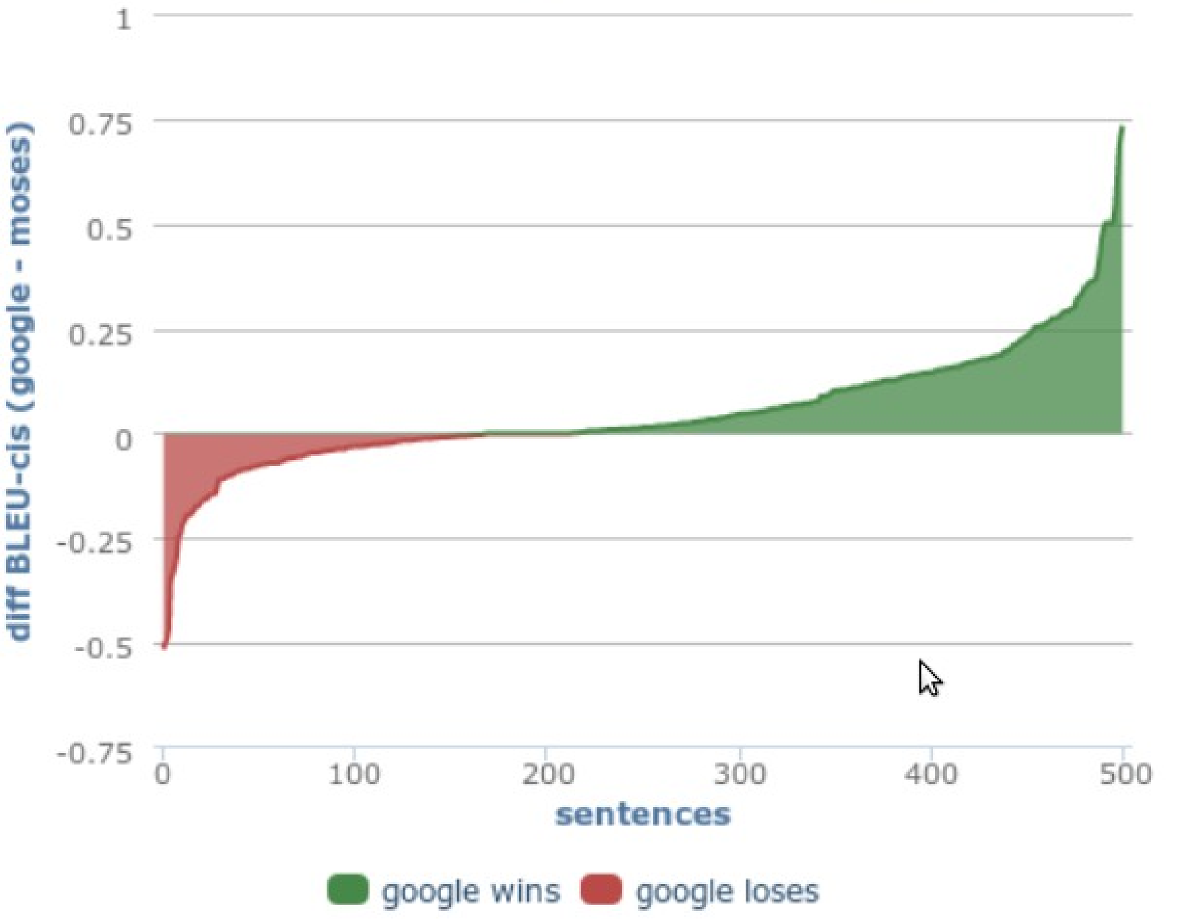
« Go back to all results

[cs] Český jazyk neboli čeština je západoslovanský jazyk, nejvíce příbuzný se slovenštinou, poté polštinou a lužickou srbštinou.

MT-ComparEval



Source	The legislators thus ignored President George Bush's appeal for them to support the plan .
Reference	Zákonodárci tak ignorovali výzvu prezidenta George Bushe , aby plán podpořili .
moses	Zákonodárci tak ignorovala výzvu prezidenta George Bushe , aby podpořil plán .
google	Zákonodárci tak ignorovali prezident George Bush odvolání pro ně podporu plánu .



Compare two MT systems

highlight and classify differences in sentences

significance tests using bootstrap resampling

Loganathan Ramasamy - Research

- **PhD Thesis**
  - Cross Lingual Annotation for Resource Poor languages
  - Preliminary Results for Tamil
- **TamilTB - Tamil Dependency Treebank**
  - Refining annotations
  - More data is planned
- **EnTam - An English-Tamil Parallel corpus**
  - Data is released
  - Approx. 170K sentences from news, cinema and Bible
- **Grants**
  - Past: CLARA
  - Present: LINDAT-CLARIN

Rudolf Rosa

Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

ÚFAL Seminar, Příchovice, 19 September 2013

Depfix (now in stand-by)

- automatic post-editing of English-to-Czech statistical machine translation outputs
- mainly rule-based corrections
  - fixing agreement errors, restoring missing negation...
- built in Treex, uses both analysis (up to t-layer) and generation (on “m-layer”)
- future plans
  - move from rule-based to machine-learnt corrections
  - extend to other languages

Khresmoi project

- translation of medical search queries for information retrieval
- handling unknown (out-of-vocabulary) words by mining synonyms from dictionary-like data (Wikipedia, MeSH...)

HamleDT (just starting)

- harmonizing treebanks for many (30) languages
  - conversion to dependencies if necessary
  - transformation to PDT-style
- past: translation to English by Google Translate
- present: adding new treebanks
- future: improving and refactoring the pipeline
- with Jan Mašek (diploma thesis)

PhD studies (just starting)

- Exploring the Structure of Natural Languages with Unsupervised and Semisupervised Methods
  - supervisor: doc. Ing. Zdeněk Žabokrtský, Ph.D.
- Can we “understand” a language unsupervisedly?
  - If not, what is the minimal supervision needed?
- What can multilingual approaches give us?
  - Can parallel information replace feedback in learning?
- How to represent language utterances?
  - Dependency trees? Other graph structures? Vectors?

Kateřina Rysová

- Participant of the GAČR

(PI: Šárka Zikánová – „Coreference, discourse relations and information structure in a contrastive perspective“)

and LINDAT (PI: Jan Hajič) grants

Kateřina Rysová

Main research interests:

- Topic-focus articulation
- Word order (esp. in Czech and German)

Current work:

- Subjective word order in Czech (based on PDT)
- Preparation of annotation of topic-focus articulation in PCEDT (in Czech and English part); together with: prof. Eva Hajičová, Jiří Mírovský, Magdaléna Rysová and annotators Olga Zitová and Klára Macháčková
- Factors influencing word order – e.g. valency (with Jiří Mírovský)

Kateřina Rysová

Further activities:

- Cooperation with Faculty of Arts: occasionally teaching, 2011–2012 investigator of GAUK project „Valency as the Word Order Factor“
- PhD-thesis: „On Word order from the Communicative Point of View“ (the defense in September 2013)
- Preparation of Olympiad in the Czech language

About me

- Shadi Saleh , Tishreen University , Latakia
- Previous projects:
  - Semester Project,2011E-Books Recommendation System,Based on Social Network profile
  - Graduation Thesis,Supervised Arabic documents classifier system

Current:

- **Working at UFAL**

  Work in PADT(Prague Arabic Dependency Treebank) project (testing scripts,filing annotation) .

- **Ph.D student**

  Cross-lingual Information retrieval Under Prof. Pavel Pecina Supervision

Milan Straka

About me

- Bc in informatics
- Mgr at Department of Algebra (cryptography)
- finishing PhD at Computer Science Institute
    - functional programming and effective data structures
- occasional work for ÚFAL since January 2012
    - Hadoop tutorial in February 2012
- started full-time in July 2013

Milan Straka

Projects

- with David Mareček – Gibbs sampler used in Unsupervised dependency parsing
- with Jana Straková – Named entity recognizer
- with Jana Straková – Log-linear classification using neural networks
- reimplementation of
  - Czech morphology
  - Czech tagger
  - Czech named entity recognizer

Teaching

- Data Intensive Computing (NPFL102) – summer 2014

Milan Straka

Reimplementation of Czech morphology

- Current implementation has not been available to public because of licensing issues
- The morphological dictionary has been recently released under ♡♡♡ CC BY-NC-SA ♡♡♡
- We will soon release new implementation which works with the released data
- C++ library with C interface, multiplatform, bindings for Perl and other needed languages

Milan Straka

Reimplementation of Czech tagger

- Originally only packaging of Featurama was planned
- A reimplementation is in progress because Featurama is too slow to be used in NER and possibly other projects
- Trained model hopefully available under CC BY-NC-SA

Reimplementation of Named entity recognizer

- Companies are interested in a Named entity recognizer
- We will soon release recognizer using the mentioned Czech morphology, Czech tagger, trained on the Czech named entity corpus
- Trained model hopefully available under CC BY-NC-SA

Jana Straková

- graduate student (”Natural language and the human brain”)
- interests:
  - natural language and the human brain
  - named entity recognition:
    - Czech Named Entity Corpus
    - Czech named entity recognizer
  - neural networks

Pavel Straňák

1. Multiword expressions, named entities
    a) MWE in PDT 2.5 ++
    b) relations btw. and inside MWEs in the dictionary
    c) NE: structure (automatic), semantics (wiki, knowl. bases, web – auto)
    d) Use for Machine translation (some MWEs already have translations …)
2. Teaching *"intro to NLP and data"* for humanities' students
3. Korektor: statistical spellchecker with bells and whistles
    a) Comparison (and combination) with MS Word spell+grammar checker
    b) Adaptation to new languages and platforms (Android, Input methods, Web)
4. LINDAT Centre, Clarin – Web Apps and Services (REST services for all apps)

Magda Ševčíková

- involved in projects
  - **GA ČR P406/12/P175** *“Selected derivational relations for automatic processing of Czech”*
    - post-doc project, 2012–2014
    - principal investigator
  - **GA ČR P406/2010/0875** *“Computational Linguistics: Explicit description of language and annotated data focused on Czech”*
    - project led by Jarmila Panevová, 2010–2013
    - team member
- teaching
  - course on academic writing *“Professional language and style”* (with Veronika Kolářová), for master students, Faculty of Mathematics and Physics
  - course on selected syntactic theories “*New directions in linguistics*”, for master students of English philology, Faculty of Philosophy and Arts
  - “*Variability of languages in time and space*” (with Anja Nedoluzhko and Šárka Zikánová), for PhD students, Faculty of Mathematics and Physics
- academic service
  - with Prof. Panevová: entrance examination tests in Czech (for applicants from abroad)

Current work:

- topics of the post-doc project:
  - deadjectival nouns with the suffix *-ost*
    - corpus-based analysis of their meanings: meaning of quality *(hloupost `stupidity')* vs. non-qualitative meaning *(neříkej hlouposti `do not say stupid things')*
    - with low-frequency words, the non-qualitative meaning indicated by formal features (plural form)
    - relation between the non-qualitative meaning and token frequency
  - productivity in word-formation
    - quantitative approaches: productivity measures based on low-frequency words in European linguistics since 1990's
    - for Czech: Miloš Dokulil's pre-corpus approach (1962) recently elaborated by František Štícha
    - pilot corpus-based study of productivity of Czech suffixes *-ost, -ství/ctví, - ita, -ismus*
  - database of Czech derivates (with Zdeněk Žabokrtský)
    - pairs of base words and their derivates
    - build-up process
      - usage of derivational information involved in morphological lemmas
      - derivation rules guessed from large corpus data
      - manually written rules
    - to be released in 2014
- revision of PDT data:
  - revision of tectogrammatical lemmas of adverbs with the suffix *-o*
  - revision of tectogrammatical annotation of negated adjectives and adverbs (t-lemma and grammateme of negation)

Jana Šindlerová

- Ph.D. Student with the topic of “contrastive study of verbal valency in Czech and English”
  - On PCEDT data
  - Intended outcome: A bilingual valency dictionary capturing alignment of verbs and verb arguments (Czengvallex) + doctoral thesis
  - Supporting grants:
    - GAUK 19008/2008 “A Multilingual Archive of Verbal Valency Characteristics” – finished
    - GPP406/13/03351P GAČR postdoc of Zdeňka Urešová: ”Srovnání české a anglické valence sloves na základě korpusového materiálu (teorie a praxe)”
    - LINDAT-Clarin

Jana Šindlerová

- Sentiment Analysis in Czech (SEANCe project) – currently with Katka Veselovská, Jan Hajič, jr., Jan Mašek (not mentioning several other kind advisors and contributors)
  - building evaluative language corpora
  - building tools for SA
  - Linguistic study of evaluative language
- Supporting grants:
  - the Grant Agency of Charles University in Prague: GAUK 353711 "Sentence-Level Polarity Detection in a Computer Corpus" – right about to finish
  - LINDAT-Clarin
  - IBM: "Sentiment Analysis extension for IBM Content Analytics"

Aleš Tamchyna

- Ph.D. student, finishing the first year.
- Research interests:
  - statistical machine translation,
  - machine learning in computational linguistics.
- Thesis topic: Lexical and Morphological Choices in Machine Translation.
- Blame me for late notices of Monday seminars!

Zdeňka Urešová I.

- **Postdoc Project (GAČR) 2013 – 2015**
  - A comparison of Czech and English verbal valency based on corpus material (theory and practice)
  - Description of verbal valency in Czech and English
  - Description of interlinking of translational verbal equivalents
  - Data preparation together with Jana Šindlerová, technical support Eva Fučíková
  - Wiki pages: https://wiki.ufal.ms.mff.cuni.cz/internal:czengvallex

- **Kreshmoi Project (EU)** with *P. Pecina, J.Hlaváčová, J. Hajič and others*
  - Topic: Medical Information on the Internet for general public and professionals
  - User Evaluation preparation for the user test cases
  - Performing the User Evaluation (May-June 2013)
  - Query translation for testing
  - Preparation of general MT test data

Zdeňka Urešová II.

- **INTLIB - Intelligent Library Project (TAČR)** *B. Hladká, V. Kříž*
  - Analytical annotation of legal texts
- **AMALACH (Ministerstvo kultury ČR – program NAKI)**
  - Localisation of the *The Visual History Archive* - online portal from USC Shoah Foundation (future, might still start in 2013)
  - Translation of a thesaurus from English to Czech (55 000 key words, in cooperation with USC)
- **LINDAT (MŠMT)**
  - Consultations for PDT-Vallex additions and editing

Zdeňka Urešová

GAČR POSTDOC PROJECT

2013 - 2015

Srovnání české a anglické valence sloves na základě korpusového materiálu (teorie a praxe)

A comparison of Czech and English verbal valency based on corpus material (theory and practice)

A Cross-linguistic Comparison of Valency Behavior of Czech and English Verbs

- **Theoretical comparative studies focused on differences in Czech and English verbal valency structure**
  - a description of verbal valency in both languages
  - a description of interlinking of translational verbal equivalents with drawing a follow-up comparison between the achieved results
  - a specification of relations of verbal valency frames in both languages, relating to PDT’s semantic and morphosyntactic levels
- **Plus hands-on experience of work with corpus data**
  - The Czech-English valency lexicon (PDT-Vallex and EngVallex) will be interlinked at the level of verb arguments, as well as linked to the data (Prague Czech-English Dependency Treebank)

1st February 2013 + 3 years

- **Goals:**
  - To describe the relation between Czech and English valency frames
  - To build a Czech-English Valency Lexicon with explicitly linked verbal senses and their arguments/adjuncts
  - A comparative description of the argument structure of translation equivalents
- **First Results:**
  - An Analysis of Annotation of Verb-Noun Idiomatic Combinations in a Parallel Dependency Corpus. The 9th Workshop on Multiword Expressions (MWE 2013), NAACL, Atlanta, Georgia, USA, June, 2013
  - Verb Valency and Argument Non-correspondence in a Bilingual Treebank. SLOVKO 2013, Bratislava, Slovakia, Nov 2013

Anna Vernerová

- advisor Markéta Lopatková
- automatic detection of applicable diatheses
  - negative phase: simple rules for excluding some <frame, diathesis> pairs (finished)
  - reflexive verbs
  - positive phase: automatically search a large corpus for instances of the derived frames (planned)
  - manual phase: providing corpus concordances for annotators who solve the undecided cases (planned)

Kateřina Veselovská

- ÚFAL since 2008
- Ph.D. student

  „Enriching the Treebank Annotation with Selected Phenomena from the Field of Pragmatics“
- in fact

  :(SEANCe:) = SEntiment ANalysis in Czech

:(SEANCe:)

GAUK 3537/2011 – *Sentence-Level Polarity Detection in a Computer Corpus*

Current team:

- Kateřina Veselovská
- Jana Šindlerová
- Jan Hajič jr.
- Jan Mašek
- supervisors: prof. Hajičová & Ondřej Bojar
- other SA people: Franky, Ondřej Fiala

GAUK 3537/2011 – *Sentence-Level Polarity Detection in a Computer Corpus*

Current state:

- sentiment-annotated corpus SubLex1.0 (4625 lemmas)
- manually annotated data from multiple domains
- several polarity classifiers with rather satisfactory results (89% accuracy)
- implementation of SubLex to TrEd (in progress)
- annotation guidelines (technical report, in progress)

Other ‘sentimental’ projects:

- sentiment analysis for IBM Content Analytics
- industrial cooperation with *Buzzboot, CaptchaWorks, Wunderman, Zoom International*
- subjectivity lexicon for Indonesian
- GAČR proposal: *On Linguistic Structure of Evaluative Meaning in Czech*

:(Kateřina Veselovská:)

Other topics of interest:

- opinion mining
- construction grammar
- tectogrammatical description of English
- parallel corpora

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

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

Dan Zeman

- Machine translation (Moses, Eman)
  - Preprocessing (word order transformations)
  - Eman (Ondřej's infrastructure)
- **Interset:** conversion of morphosyntactic tags between tagsets (both in Czech and cross-language)
  - Universal description of morphological tagsets
- **HamleDT:** Multilingual dependency parsing
  - With Zdeněk, MartinP, David, JanŠ, Loganathan…

Dan Zeman

- Parsing
  - Nivre’s Malt Parser on Czech (UAS 86.0 % on d-test and 85.8 % on e-test)
- Teaching
  - Morphological and Syntactic Analysis
  - Disrupted: Computational NLP (but kept at ČVUT)
  - New course: “New Language”
- “Dirty” (non-scientific) work
  - Bibliography maintenance (the “Biblio” database)
  - Address book maintenance (PBML, corpora registration…)

GAČR P406/11/1499

- 2011 – 2013
- Titled *Czech in the Machine Translation Era (CZECHMATE)*
    - Dan Zeman
    - Ondřej Bojar
- Non-English translation (e.g. Czech-German)
- Phrase-based translation
- Named entities

Coreference, discourse relations and information structure in a contrastive perspective

GA ČR P406/12/0658

PI: Šárka Zikánová

- Standard GA ČR project 2012-15

- Prof. Hajičová, Pavlína Jínová, Jiří Mírovský, Anja Nedoluzhko, Lucie Poláková, Kateřina Rysová, Magdaléna Rysová, Barbora Vidová Hladká, Šárka Zikánová

Discourse

- Finishing the discourse annotation in the PDT, publication of the first version (2012, PDiT)
  - Explicit traditional discourse connectives and the adjacent discourse arguments; intra- and inter-sentential relations

- Current work on the second version which will be included in the PDT 3.0
  - Genre annotation (see the talk by Lucie Poláková)

- Alternative lexicalizations (Magdaléna Rysová)

Information structure

- The ordering of contextually non-bound nodes depending on a verb (Kateřina Rysová, PhD thesis, see her talk)
  - Influences of valency, semantic richness, form of the expression
  - Comparison with German
- Definite and indefinite NPs in English and their counterparts in Czech (prof. Hajičová, Jiří Mírovský)
- Acquisition of salience diagrams from the existing annotation in the PDT (prof. Hajičová, Barbora Vidová Hladká)

Coreference

- Annotation of coreference of the 1st and 2nd persons for the PDT 3.0 (Anja Nedoluzhko, Jiří Mírovský)
- Annotation and research of places without coreference or discourse relations (Šárka Zikánová)

Further plans

- Interplays: information structure, discourse, coreference
  - Indefiniteness, pronominalization and ellipsis
  - Segmentation of discourse in relation to the coreference
  - …
- The contrastive perspective: Czech, English, German
  - Manual for annotation of the information structure in English and comparison with Czech
  - Comparative studies on single discourse connectives and on sets of discourse relations (are they universal?)
  - Subjective word order in Czech, English and German
  - …

Zdeněk Žabokrtský (1/4)

- **topics of interest**
  - **past**
    - valency frames of verbs (VALLEX)
    - treebanking (PDT)
    - anaphora resolution, parsing, named entities
  - **current**
    - dependency syntax in applications (Treex)
    - dependency syntax accross languages (HamleDT)
    - machine learning (only a modest consumer of)
    - building a word formation network for Czech

Zdeněk Žabokrtský (2/4)

- **office**
  - **organizing PhD events**
    - PhD defenses
    - state doctoral exams
  - **ÚFAL internships**
    - up to 3-month research opportunities for foreign students
    - one candidate selected twice or three times a year

Zdeněk Žabokrtský (3/4)

- **Courses in 2013/2014:**
  - **Technology for NLP**
    - bash+perl+xml...
  - **Language data resources** (with Martin Popel)
    - corpora, treebanks, lexical databases …
  - **Selected Problems in Machine Learning** (with David Mareček)
    - intro to Bayesian ML, Gibbs sampling..., for PGS
  - **Exercises in Machine Learning** (with Ondřej Bojar)
    - gaining experience on various ML techniques
  - **Variability of Languages in Time and Space**

    (taught by Magda Ševčíková, Šárka Zikánová and Anja Nedoluzhko)
    - serving only as a technical support for exercises

Zdeněk Žabokrtský (4/4)

- **Students supervised in 2013/2014:**
  - **PhD students**
    - Martin Popel
    - Loganthan Ramasamy
    - Michal Novák
    - Rudolf Rosa
  - **Master students**
    - Jan Mašek
    - Václav Honzík (ČVUT)