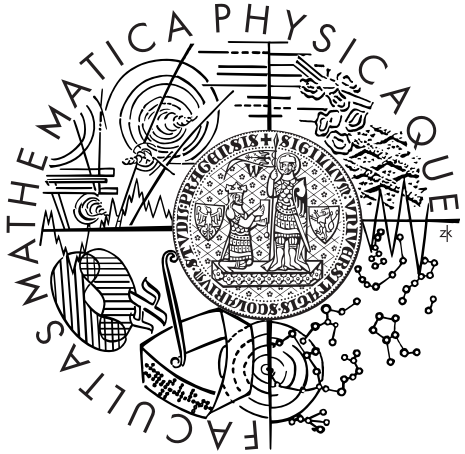
Short presentations



ÚFAL seminar
Sedlec-Prčice
13. – 16. 9. 2014

Last update: 15. 9. 2014 at 22:58

Contents

Lukáš Žilka

Nic než piliny

Šárka Zikánová

Projects

- GA ČR: standard project

  Coreference, discourse relations and information structure in a contrastive perspective (P406/12/0658)

- KONTAKT: cooperation with University of Pennsylvania

  Multilingual corpus annotation as a support for language technologies (KONTAKT 14011)

Teaching

Courses:

- Information structure of sentences and discourse structure (with E. Hajičová)
- Variability of languages in time and space (with M. Ševčíková and A. Nedoluzhko)

- Supervisor of Ph.D. theses and master theses

Coreference, discourse relations and information structure in a contrastive perspective

GA ČR P406/12/0658

PI: Šárka Zikánová

ÚFAL

- 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

- 3 basic topics:
  - Discourse
  - Coreference
  - Information structure

Discourse

- Representation of Prague discourse conception to the Czech academic community (submitted for publictaion in Slovo a slovesnost, terminology)
- Annotation of alternative lexicalizations of discourse connectives (*the reason is* instead of *because*) finished; checking consistency of annotation in progress, annotators´ agreement measured (so far preliminary results)
- Automatic searching for multiword discourse markers
- Distribution of discourse types across genres measured and described

Information structure

- Monograph "O slovosledu z komunikačního pohledu" (The word order from the communicative perspective) – Ph.D. thesis by Kateřina Rysová (Bolzano award)
- Influence of valency on the word order in Czech (LREC 2014, K. Rysová – J. Mírovský)
- Automatic annotation of the topic-focus articulation in the PCEDT (with J. Mírovský)

Coreference

- Comparison of coreferential expressions in Czech and English (Anja Nedoluzhko with M. Novák)
- Just starting: comparison of approaches to discourse and cohesive phenomena in Czech, English and German (with colleagues in Saarbrucken)
- Annotation of nominal coreference annotation in PCEDT (ready, in Czech part some controls remaining)

Interplays

- Preparation of a collective monogtraph “Text structure: the interplay of the information structure, coreference and discourse relations” (to be submitted in July 2015, for the UFAL book series)

Multilingual Corpus Annotation as a Support for Language Technologies KONTAKT 14011

Šárka Zikánová

KONTAKT 14011

- 2014 - 2016
- Partner: University of Pennsylvania, Institute for Research in Cognitive Science

  Aravind K. Joshi, Bonnie Webber, Rashmi Prasad (Penn Discourse Treebank)

- Comparative description of Czech and English text structure
- Deduction of the meaning of a text
  - Discourse
  - Coreference

Cooperation

- Long informal cooperation
- KONTAKT project *Towards a computational analysis of text structure* (2010 – 2012)
- New know-how for both sides, theoretical feedback, feedback coming from the data of a different language
- Plans:
  - Text genres
  - Prediction of implicit discourse relations

Dan Zeman

- Funding
  - PRVOUK
  - KHRESMOI => Statistical Machine Translation
    - ended in August
  - QT LEAP
    - since September

Dan Zeman

- **Interset:** conversion of morphosyntactic tags between tagsets (both in Czech and cross-language)
  - Universal description of morphological tagsets
- **HamleDT**
  many treebanks → common annotation style
  - With Martin P., Zdeněk, Ruda and others
- **Universal Dependencies**
  - With Joakim Nivre, Google folks and others

Dan Zeman

- Parsing
  - Underresourced languages, delexicalized parsing
- Teaching
  - Morphological and Syntactic Analysis
  - Disrupted: Computational NLP (but kept at ČVUT)
  - New course: “New Language”
- Other
  - Bibliography maintenance (the “Biblio” database)

Anna Vernerová

Advisor: Markéta Lopatková

- annotation of diatheses for Vallex
  - finished:
    - the recipient diathesis
    - the possessive resultative diathesis
  - ongoing:
    - the passive diathesis
- nominal entries for Vallex

Zdeňka Urešová

- **GAČR POSTDOC PROJECT** (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
  – Data preparation together with Jana Šindlerová, Eva Fučíková
  – Related: AMR annotation comparison between Czech and English (par t of JHU Workshop CLAMR team, Kontakt II)
  – Publ in 2014: 4 (Parseme, LREC, COLING workshop LG-LP, ACL Workshop on "Events")

- **EU KHRESMOI PROJECT** (2010-2014, finished - Aug. 2014)

  Medical Information on the Internet for general public and professionals (PI: J. Hajič)

  – User Evaluation (together with J. Hlaváčová and J.Hajič)
  – Query translation for testing (mostly medical terms)
  – Preparation of general MT test data
  – Publ in 2014: 7 with others in the project (see http://khresmoi.eu)

- **MK "NAKI": AMALACH** (Ministry of Culture, Czech Rep., 2012-2015)

  General goal: cross-language indexing of audio in Czech and English, with Univ. of West Bohemia

  – Localisation of the The Visual History Archive online portal from USC Shoah Foundation
  – Translation of a thesaurus from English to Czech (55 000 keywords, in cooperation with USC)

Zdeňka Urešová

- **SERVICE AND OTHER ACTIVITIES**
  - **Depling 2013, Prague** – poster session organization
  - **JHU Workshop 2014, Prague** – organization + member of CLAMR (=CLAMR: Cross-Lingual Abstract Meaning Representations for Machine Translation; led by Martha Palmer
    - Main task: annotation of AMR for Czech texts
  - **Reviewer** for Program Committee, Area "Resources", **COLING 2014**, Dublin, Ireland (see http://www.coling-2014.org)
  - **Member of Program** Committee for 10th Joint ACL - ISO **Workshop on Interoperable Semantic Annotation**, Reykjavik, Iceland, May 26, 2014, see http://sigsem.uvt.nl/isa10/
  - Co-author of **ISO 24617-4:2014** (en)
    - Language resource management — Semantic annotation framework (SemAF)
    - Part 4: Semantic roles (SemAF-SR)
  - **Talk:**
    - Czech lexical resources for NLP – Google NYC, Feb. 2014

Aleš Tamchyna

- PhD student at ÚFAL
  - advised by Ondřej Bojar
  - starting my 3rd year
- research interests:
  - statistical machine translation
  - machine learning in NLP

Magda Ševčíková

- involved in projects
  - GA ČR P406/12/P175 ***Selected derivational relations for automatic processing of Czech***
    - principal investigator, post-doc project, 2012–2014
  - LM2010013 **LINDAT-Clarin**
    - team member
  - (proposal of a GA ČR project on description of morphological and syntactic phenomena based on Prague dependency treebanks of spoken and written Czech, submitted for 2015–2017)
- teaching
  - course on academic writing *"Professional language and style"*
    - with Veronika Kolářová
    - for master students, Faculty of Mathematics and Physics
  - *"Introduction to Formal Linguistics"*
    - 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
    - not taught in Fall 2014
- academic service
  - entrance examination tests in Czech (for applicants from abroad)
    - with Prof. Panevová

Current work:

- database of Czech derived words “DeriNet”
  - with Zdeněk Žabokrtský
  - DeriNet version 0.5 released in May 2014 at `http://ufal.mff.cuni.cz/derinet`
  - version 1.0 expected by December 2014
  - more than 260k noun, adjective, verb, and adverb lemmas from the SYN subcorpus of the Czech National Corpus
  - lemmas (nodes) connected with links (edges) corresponding to derivational relations
    - 46k derivational links delivered by an existing tool for morphological analysis (Hajič 2004)
    - 26k links based on an automatically discovered set of derivation rules
    - 4k links created by a (manually written) grammar-based set of derivation rules
    - ... under construction
  - presented at
    - LREC 2014, May 26–31, Reykjavik, Iceland
    - SLE 2014 (Societas Linguistica Europaea), September 11–14, Poznan, Poland
    - (at a Monday seminar in Fall 2014)
- linguistic research in word-formation
  - semantics and frequency characteristics of the Czech noun suffix *-ost*
    - published in journal *Korpus – gramatika – axiologie* 9, 2014
  - productivity od nouns with suffixes *-ost, -ství/-tví, -ismus, -ita* (meaning of quality)
    - accepted for publication in journal *Naše řeč* in 2014
  - revision of grammatical description of the suffix *-ství* – a case study based on corpus and database data
    - to be presented at the Corpus Linguistics Conference in Prague, September 17–19

Pavel Straňák

Projects

1. LINDAT/CLARIN “scientific secretary”
    - Web Apps and Services, repository
    - Korektor: statistical spellchecker with bells and whistles; some plans with Milan Straka (starting with training for 1-2 other languages)
    - anything else
2. EUDAT (B2SHARE): a big part of the past year (finished)
3. PARSEME: Multiword expressions, named entities
    - MWE in PDT 3 ++
    - Named Entities: structure (automatic), semantics (wiki, knowl. bases, web – auto)
    - Parsing of MWEs
    - Use of MWE-specific information in other applications

Teaching

NPFL098 (ATKL00345): Automatické zpracování textových dat

- The goal is to facilitate smooth transition to UFAL courses that require some programming and text (data) processing skills.
- for humanities' students or our doctoral students without background in computer science
  - basic unix + perl data processing, Unicode, using REST web services, OpenNLP models training …
  - since 2013/14 extended to 2+2

Common Good

ISBN evidence: I assign UFAL ISBNs to datasets or books

Jana Straková

- graduate student (advisor: prof. Jan Hajič)
- research interests:
  - cognitive neuroscience
  - artificial neural networks
  - tagging, named entity recognition
- projects:
  - MorphoDiTa
  - NameTag
  - Czech Named Entity Corpus

Milan Straka

About me

- Bc in informatics
- Mgr at Department of Algebra (cryptography)
- PhD at Department of Applied Mathematics
  - functional programming and effective data structures
- occasional work for ÚFAL since January 2012
  - Hadoop tutorial in February 2012
  - Gibbs sampler implementation used by David Mareček in Unsupervised dependency parsing in Autumn 2012
- started full-time in July 2013

Milan Straka

About me

- Bc in informatics
- Mgr at Department of Algebra (cryptography)
- PhD at Department of Applied Mathematics
  - functional programming and effective data structures
- occasional work for ÚFAL since January 2012
  - Hadoop tutorial in February 2012
  - Gibbs sampler implementation used by David Mareček in Unsupervised dependency parsing in Autumn 2012
- started full-time in July 2013

Milan Straka

About me

- Bc in informatics
- Mgr at Department of Algebra (cryptography)
- PhD at Department of Applied Mathematics
  - functional programming and effective data structures
- occasional work for ÚFAL since January 2012
  - Hadoop tutorial in February 2012
  - Gibbs sampler implementation used by David Mareček in Unsupervised dependency parsing in Autumn 2012
- started full-time in July 2013

Milan Straka

About me

- Bc in informatics
- Mgr at Department of Algebra (cryptography)
- PhD at Department of Applied Mathematics
  - functional programming and effective data structures
- occasional work for ÚFAL since January 2012
  - Hadoop tutorial in February 2012
  - Gibbs sampler implementation used by David Mareček in Unsupervised dependency parsing in Autumn 2012
- started full-time in July 2013

Milan Straka

About me

- Bc in informatics
- Mgr at Department of Algebra (cryptography)
- PhD at Department of Applied Mathematics
  - functional programming and effective data structures
- occasional work for ÚFAL since January 2012
  - Hadoop tutorial in February 2012
  - Gibbs sampler implementation used by David Mareček in Unsupervised dependency parsing in Autumn 2012
- started full-time in July 2013

Milan Straka

About me

- Bc in informatics
- Mgr at Department of Algebra (cryptography)
- PhD at Department of Applied Mathematics
  - functional programming and effective data structures
- occasional work for ÚFAL since January 2012
  - Hadoop tutorial in February 2012
  - Gibbs sampler implementation used by David Mareček in Unsupervised dependency parsing in Autumn 2012
- started full-time in July 2013

Projects

- MorphoDiTa – **Morpho**logical **Di**ctionary and **Ta**gger
  - implementation of morphological dictionary and POS tagger, performing morphological analysis, morphological generation, POS+lemma tagging and UTF-8 tokenization
  - available models for Czech and English under CC BY-NC-SA license
  - Slovak, Swedish and hopefully Arabic coming soon
    - *any favourite languages you need models for?*
  - multiplatform (Linux/Windows/OS X) C++ implementation under LGPL, bindings for Java, Perl (CPAN) and Python (PyPI)
  - available also as LINDAT/CLARIN service, Weblicht integration coming soon

Projects

- MorphoDiTa – **Morpho**logical **Di**ctionary and **Ta**gger
  - implementation of morphological dictionary and POS tagger, performing morphological analysis, morphological generation, POS+lemma tagging and UTF-8 tokenization
  - available models for Czech and English under CC BY-NC-SA license
  - Slovak, Swedish and hopefully Arabic coming soon
    - *any favourite languages you need models for?*
  - multiplatform (Linux/Windows/OS X) C++ implementation under LGPL, bindings for Java, Perl (CPAN) and Python (PyPI)
  - available also as LINDAT/CLARIN service, Weblicht integration coming soon

Projects

- MorphoDiTa – **Morpho**logical **Di**ctionary and **Ta**gger
  - implementation of morphological dictionary and POS tagger, performing morphological analysis, morphological generation, POS+lemma tagging and UTF-8 tokenization
  - available models for Czech and English under CC BY-NC-SA license
  - Slovak, Swedish and hopefully Arabic coming soon
    - *any favourite languages you need models for?*
  - multiplatform (Linux/Windows/OS X) C++ implementation under LGPL, bindings for Java, Perl (CPAN) and Python (PyPI)
  - available also as LINDAT/CLARIN service, Weblicht integration coming soon

Projects

- MorphoDiTa – **Morpho**logical **Di**ctionary and **Ta**gger
  - implementation of morphological dictionary and POS tagger, performing morphological analysis, morphological generation, POS+lemma tagging and UTF-8 tokenization
  - available models for Czech and English under CC BY-NC-SA license
  - Slovak, Swedish and hopefully Arabic coming soon
    - *any favourite languages you need models for?*
  - multiplatform (Linux/Windows/OS X) C++ implementation under LGPL, bindings for Java, Perl (CPAN) and Python (PyPI)
  - available also as LINDAT/CLARIN service, Weblicht integration coming soon

Projects

- MorphoDiTa – **Morpho**logical **Di**ctionary and **Ta**gger
  - implementation of morphological dictionary and POS tagger, performing morphological analysis, morphological generation, POS+lemma tagging and UTF-8 tokenization
  - available models for Czech and English under CC BY-NC-SA license
  - Slovak, Swedish and hopefully Arabic coming soon
    - *any favourite languages you need models for?*
  - multiplatform (Linux/Windows/OS X) C++ implementation under LGPL, bindings for Java, Perl (CPAN) and Python (PyPI)
  - available also as LINDAT/CLARIN service, Weblicht integration coming soon

Projects

- MorphoDiTa – **Morpho**logical **Di**ctionary and **Ta**gger
  - implementation of morphological dictionary and POS tagger, performing morphological analysis, morphological generation, POS+lemma tagging and UTF-8 tokenization
  - available models for Czech and English under CC BY-NC-SA license
  - Slovak, Swedish and hopefully Arabic coming soon
    - *any favourite languages you need models for?*
  - multiplatform (Linux/Windows/OS X) C++ implementation under LGPL, bindings for Java, Perl (CPAN) and Python (PyPI)
  - available also as LINDAT/CLARIN service, Weblicht integration coming soon

Projects

- MorphoDiTa – **Morpho**logical **Di**ctionary and **Ta**gger
  - implementation of morphological dictionary and POS tagger, performing morphological analysis, morphological generation, POS+lemma tagging and UTF-8 tokenization
  - available models for Czech and English under CC BY-NC-SA license
  - Slovak, Swedish and hopefully Arabic coming soon
    - *any favourite languages you need models for?*
  - multiplatform (Linux/Windows/OS X) C++ implementation under LGPL, bindings for Java, Perl (CPAN) and Python (PyPI)
  - available also as LINDAT/CLARIN service, Weblicht integration coming soon

Projects

- NameTag – **Name**d entity **tag**ger
  - named entity recognizer build upon MorphoDiTa
  - available models for Czech under CC BY-NC-SA license, internal models for English (hopefully to be released soon)
  - once again multiplatform C++ implementation under LGPL with Java, Perl (CPAN) and Python (PyPI) binding
  - available also as LINDAT/CLARIN service, Weblicht integration coming soon
- CNEC 2.0 – **C**zech **N**amed **E**ntity **C**orpus 2.0
- with Pavel Straňák – improving Korektor
  - small model improvements and bug fixes
  - multithreaded LINDAT/CLARIN service, Weblicht to come
  - web browser integration planned

Projects

- NameTag – **Name**d entity **tag**ger
  - named entity recognizer build upon MorphoDiTa
  - available models for Czech under CC BY-NC-SA license, internal models for English (hopefully to be released soon)
  - once again multiplatform C++ implementation under LGPL with Java, Perl (CPAN) and Python (PyPI) binding
  - available also as LINDAT/CLARIN service, Weblicht integration coming soon
- CNEC 2.0 – **C**zech **N**amed **E**ntity **C**orpus 2.0
- with Pavel Straňák – improving Korektor
  - small model improvements and bug fixes
  - multithreaded LINDAT/CLARIN service, Weblicht to come
  - web browser integration planned

Milan Straka

Projects

- NameTag – **Name**d entity **tag**ger
  - named entity recognizer build upon MorphoDiTa
  - available models for Czech under CC BY-NC-SA license, internal models for English (hopefully to be released soon)
  - once again multiplatform C++ implementation under LGPL with Java, Perl (CPAN) and Python (PyPI) binding
  - available also as LINDAT/CLARIN service, Weblicht integration coming soon
- CNEC 2.0 – **C**zech **N**amed **E**ntity **C**orpus 2.0
- with Pavel Straňák – improving Korektor
  - small model improvements and bug fixes
  - multithreaded LINDAT/CLARIN service, Weblicht to come
  - web browser integration planned

Projects

- NameTag – **Name**d entity **tag**ger
  - named entity recognizer build upon MorphoDiTa
  - available models for Czech under CC BY-NC-SA license, internal models for English (hopefully to be released soon)
  - once again multiplatform C++ implementation under LGPL with Java, Perl (CPAN) and Python (PyPI) binding
  - available also as LINDAT/CLARIN service, Weblicht integration coming soon
- CNEC 2.0 – **C**zech **N**amed **E**ntity **C**orpus 2.0
- with Pavel Straňák – improving Korektor
  - small model improvements and bug fixes
  - multithreaded LINDAT/CLARIN service, Weblicht to come
  - web browser integration planned

Projects

- NameTag – **Name**d entity **tag**ger
  - named entity recognizer build upon MorphoDiTa
  - available models for Czech under CC BY-NC-SA license, internal models for English (hopefully to be released soon)
  - once again multiplatform C++ implementation under LGPL with Java, Perl (CPAN) and Python (PyPI) binding
  - available also as LINDAT/CLARIN service, Weblicht integration coming soon
- CNEC 2.0 – Czech Named Entity Corpus 2.0
- with Pavel Straňák – improving Korektor
  - small model improvements and bug fixes
  - multithreaded LINDAT/CLARIN service, Weblicht to come
  - web browser integration planned

Projects

- NameTag – **Name**d entity **tag**ger
  - named entity recognizer build upon MorphoDiTa
  - available models for Czech under CC BY-NC-SA license, internal models for English (hopefully to be released soon)
  - once again multiplatform C++ implementation under LGPL with Java, Perl (CPAN) and Python (PyPI) binding
  - available also as LINDAT/CLARIN service, Weblicht integration coming soon
- CNEC 2.0 – **C**zech **N**amed **E**ntity **C**orpus 2.0
- with Pavel Straňák – improving Korektor
  - small model improvements and bug fixes
  - multithreaded LINDAT/CLARIN service, Weblicht to come
  - web browser integration planned

Milan Straka

Projects

- NameTag – **Name**d entity **tag**ger
  - named entity recognizer build upon MorphoDiTa
  - available models for Czech under CC BY-NC-SA license, internal models for English (hopefully to be released soon)
  - once again multiplatform C++ implementation under LGPL with Java, Perl (CPAN) and Python (PyPI) binding
  - available also as LINDAT/CLARIN service, Weblicht integration coming soon
- CNEC 2.0 – **C**zech **N**amed **E**ntity **C**orpus 2.0
- with Pavel Straňák – improving Korektor
  - small model improvements and bug fixes
  - multithreaded LINDAT/CLARIN service, Weblicht to come
  - web browser integration planned

Milan Straka

Projects

- NameTag – **Name**d entity **tag**ger
  - named entity recognizer build upon MorphoDiTa
  - available models for Czech under CC BY-NC-SA license, internal models for English (hopefully to be released soon)
  - once again multiplatform C++ implementation under LGPL with Java, Perl (CPAN) and Python (PyPI) binding
  - available also as LINDAT/CLARIN service, Weblicht integration coming soon
- CNEC 2.0 – **C**zech **N**amed **E**ntity **C**orpus 2.0
- with Pavel Straňák – improving Korektor
  - small model improvements and bug fixes
  - multithreaded LINDAT/CLARIN service, Weblicht to come
  - web browser integration planned

Milan Straka

Projects

- NameTag – **Name**d entity **tag**ger
  - named entity recognizer build upon MorphoDiTa
  - available models for Czech under CC BY-NC-SA license, internal models for English (hopefully to be released soon)
  - once again multiplatform C++ implementation under LGPL with Java, Perl (CPAN) and Python (PyPI) binding
  - available also as LINDAT/CLARIN service, Weblicht integration coming soon
- CNEC 2.0 – **C**zech **N**amed **E**ntity **C**orpus 2.0
- with Pavel Straňák – improving Korektor
  - small model improvements and bug fixes
  - multithreaded LINDAT/CLARIN service, Weblicht to come
  - web browser integration planned

Milan Straka

Projects

- NameTag – **Name**d entity **tag**ger
  - named entity recognizer build upon MorphoDiTa
  - available models for Czech under CC BY-NC-SA license, internal models for English (hopefully to be released soon)
  - once again multiplatform C++ implementation under LGPL with Java, Perl (CPAN) and Python (PyPI) binding
  - available also as LINDAT/CLARIN service, Weblicht integration coming soon
- CNEC 2.0 – **C**zech **N**amed **E**ntity **C**orpus 2.0
- with Pavel Straňák – improving Korektor
  - small model improvements and bug fixes
  - multithreaded LINDAT/CLARIN service, Weblicht to come
  - web browser integration planned

Milan Straka

Research

- with Jana Straková – neural network classifiers
  - NameTag uses simple neural network classifier, outperforms classifiers we compared it to (mostly maximum entropy classifiers)
    - nevertheless, we did not compare to CRF on same data
  - hoping to achieve state-of-the-art results also in other areas

Teaching

- Data Intensive Computing (NPFL102) – summer 2014
  - distributed computations using SGE/OGE, Hadoop, Spark and PySpark
  - Spark will be available on ÚFAL cluster coon

Milan Straka

Research

- with Jana Straková – neural network classifiers
  - NameTag uses simple neural network classifier, outperforms classifiers we compared it to (mostly maximum entropy classifiers)
    - nevertheless, we did not compare to CRF on same data
  - hoping to achieve state-of-the-art results also in other areas

Teaching

- Data Intensive Computing (NPFL102) – summer 2014
  - distributed computations using SGE/OGE, Hadoop, Spark and PySpark
  - Spark will be available on ÚFAL cluster coon

Milan Straka

Research

- with Jana Straková – neural network classifiers
  - NameTag uses simple neural network classifier, outperforms classifiers we compared it to (mostly maximum entropy classifiers)
    - nevertheless, we did not compare to CRF on same data
  - hoping to achieve state-of-the-art results also in other areas

Teaching

- Data Intensive Computing (NPFL102) – summer 2014
  - distributed computations using SGE/OGE, Hadoop, Spark and PySpark
  - Spark will be available on ÚFAL cluster coon

Milan Straka

Research

- with Jana Straková – neural network classifiers
  - NameTag uses simple neural network classifier, outperforms classifiers we compared it to (mostly maximum entropy classifiers)
    - nevertheless, we did not compare to CRF on same data
  - hoping to achieve state-of-the-art results also in other areas

Teaching

- Data Intensive Computing (NPFL102) – summer 2014
  - distributed computations using SGE/OGE, Hadoop, Spark and PySpark
  - Spark will be available on ÚFAL cluster coon

Milan Straka

Research

- with Jana Straková – neural network classifiers
  - NameTag uses simple neural network classifier, outperforms classifiers we compared it to (mostly maximum entropy classifiers)
    - nevertheless, we did not compare to CRF on same data
  - hoping to achieve state-of-the-art results also in other areas

Teaching

- Data Intensive Computing (NPFL102) – summer 2014
  - distributed computations using SGE/OGE, Hadoop, Spark and PySpark
  - Spark will be available on ÚFAL cluster coon

Milan Straka

Research

- with Jana Straková – neural network classifiers
  - NameTag uses simple neural network classifier, outperforms classifiers we compared it to (mostly maximum entropy classifiers)
    - nevertheless, we did not compare to CRF on same data
  - hoping to achieve state-of-the-art results also in other areas

Teaching

- Data Intensive Computing (NPFL102) – summer 2014
  - distributed computations using SGE/OGE, Hadoop, Spark and PySpark
  - Spark will be available on ÚFAL cluster coon

Magdaléna Rysová

A team member of grants and projects:

GAČR

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

Kontakt II

PI: Šárka Zikánová – Multilingual Corpus Annotation as a Support for Language Technologies

Cost

Czech PI: Jiří Mírovský – Structuring Discourse in Multilingual Europe (TextLink)

Magdaléna Rysová

Main Research Interests

- Discourse relations
- Discourse connectives and other means of expressing textual relations
- Theme of Ph.D. thesis: Discourse connectives expressed by multiword expressions like „the reason is“, „this means“, „the condition is“ etc. (supervisor: prof. PhDr. Eva Hajičová, DrSc.)

Kateřina Rysová

Participant of following grants and projects:

- GAČR

PI: Šárka Zikánová – *Coreference, discourse relations and information structure in a contrastive perspective*

- Kontakt II

PI: Šárka Zikánová – *Multilingual Corpus Annotation as a Support for Language Technologies*

- Cost

Czech PI: Jiří Mírovský – *Structuring Discourse in Multilingual Europe* (*TextLink*)

- LINDAT

PI: Jan Hajič

Kateřina Rysová

Main research interests:

- Topic-focus articulation
- Word order
- Discourse analysis

Current work:

Preparation of the monograph *On Word Order from the Communicative Point of View*

Preparation of annotation of topic-focus articulation in PCEDT (in Czech and English part); together esp. with: prof. Eva Hajičová, Jiří Mírovský

Kateřina Rysová

Further activities:

- Cooperation with Faculty of Arts: occasionally teaching, investigator of the finished GAUK project *Valency as the Word Order Factor*
- Preparation of Olympiad in the Czech language

Rudolf Rosa

- PhD student of Zdeněk Žabokrtský (1st year)
  - parsing in multilingual setting, semi-supervised methods
- GAUK (with Jan Mašek, supervised by ZŽ)
  - modelling dependency syntax across languages
- HamleDT project
  - harmonizing annotation of dependency treebanks
- QTLeap grant
  - machine translation with deep language processing
  - Depfix – automatic post-editing of machine translation

Loganathan Ramasamy - Research

- **PhD Thesis**
  - Submitted
  - Thesis: "Parsing under-resourced languages: Cross-lingual transfer strategies for Indian languages"
- **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
- **Publications - 2014**
  - To appear in PBML: "Multilingual Dependency Parsing: Using Machine Translated Texts instead of Parallel Corpora"
- **Grants**
  - Past: CLARA
  - Present: LINDAT/CLARIN

Martin Popel

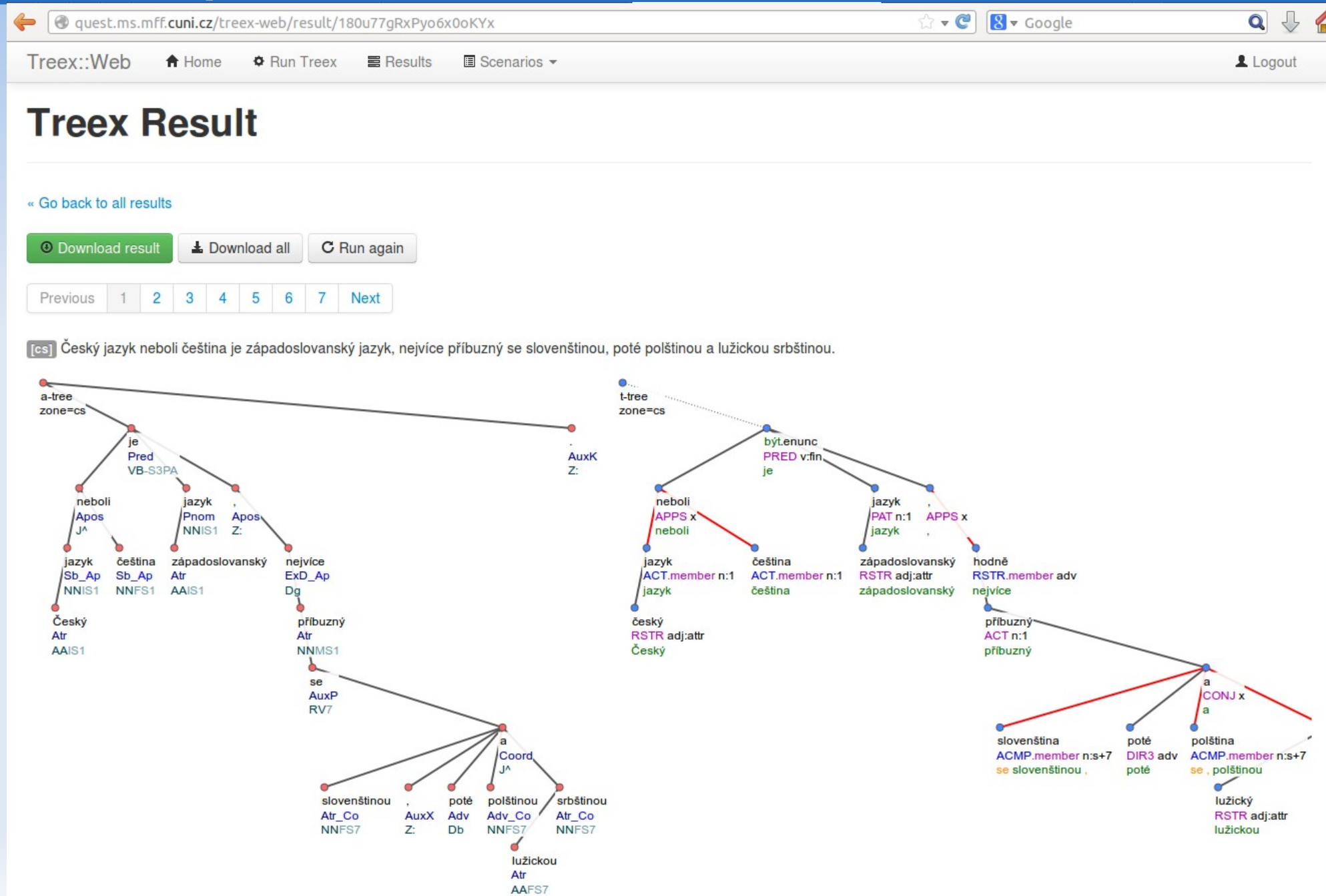
- **Treex** NLP framework (+ Treex::Web by Michal Sedlák)
  - more language-independent code for tecto-analysis (and synthesis) exploiting Interset (by Dan Zeman)
- **TectoMT** machine translation (PhD thesis on transfer)
  - experiments with Vowpal Wabbit and structured prediction
- **HamleDT** 30+ treebanks (ACL14: stanfordization, LRE paper)
- **PBML** (next deadline: January 15th 2015) LaTeX volunteer?
- **Technical reports** (2014 deadline: December 1st)
- **Teaching** autumn: Modern Methods in CL I ("Reading group")

  spring: Modern Methods in CL II (for PhD and staff, Deep NN ?)

  Language Data Resources (with ZŽ, me: significance)

Treex::Web

https://lindat.mff.cuni.cz/services/treex-web/



QTLeap

Quality Translation by Deep Language Engineering Approaches

- 8 partners: (FCUL, DFKI, CUNI, IICT-BAS, UBER, UPV/EHU, UG, HF)
- Build MT system(s) for 14 language pairs:
  English ↔ Basque, Bulgarian, Czech, Dutch, German, Portuguese, Spanish
- Extrinsic evaluation: IT-related question answering
  (user questions translated to English → IR → answers translated back)
- TectoMT considered for MT Pilot 1 for:
  Czech, Dutch, Portuguese, Spanish
- Depfix considered for: Basque (and Czech via Chimera)
- ÚFAL leads WP2 ”Deep MT” (55 PMs)
- 14 PMs also on WP5 ”Lexical Semantics: Linking and resolving” (Linked Open Data)

Lucie Poláková

Project:

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

Recently:

- PDiT 1.0 released in November 2012
- Enhanced version released in PDT 3.0 in December 2013 (new: mainly genres, rhematizers, Anja – coreference of 1st + 2nd person)
- Annotation of alternative lexicalizations od discourse connective in PDT (Majda Rysová)

Grant support:

GAČR (Zikánová, "Interplays" till 2015)

LINDAT

KONTAKT, second round - Cooperation with UPenn and prof. Aravind Joshi's team (+ dr. R. Prasad in Milwaukee, Winconsin)

Lucie Poláková

COST: TextLink (Action IS1312)

Huge EU project on multilingual corpora with linked connectives – funding mainly for gathering scientists and resources, no salaries (PI – L. Degand, Belgium)

MC members for Czech Republic: Jiří Mírovský, L. Poláková

April 2014: kick-off meeting in Brussels

October 20.-21. 2014 **meeting in Prague** (L. Degand, B. Webber, M. Stede etc.)

Note: **STSM** - **Short term scientific missions – call for participation**, scientific stays from the COST funding in the whole Europe (+ Turkey)

Lucie Poláková

PhD: Dissertation topic:

The concept of discourse-level description for the PDT

Administrative:

- new UFAL webpages
- nástěnkářka ☺

(poster decoration art manager)

Michal Novák

- GAUK 4226/2011: Utilization of coreference in machine translation
  - finished
- GAČR project proposal: Anaphora resolution in Czech and its linguistic analysis
  - with Anja, Šárka, Jirka and Ivana
- coreference-related topics:
  - Anaphoric expressions in parallel Czech-English data
    - with Anja
  - Cross-lingual coreference resolution
    - with Zdeněk
  - semi-supervised approaches for cross-lingual CR

Michal Novák

- Khresmoi
  - finished
  - data filtering for MT
  - Czech NLP pipeline for information extraction within the Gate toolkit
- public service
  - supplying Karolinum bookstore with books published at ÚFAL
  - administration of the related web pages

Sales of ÚFAL books

Book	11/12 – 03/13	04/13 – 03/14	04/14 – 06/14	Total
Ondřej Bojar: Čeština a strojový překlad: Strojový překlad našincům, našinci strojovému překladu	7	2		9
Ondřej Bojar: Exploiting linguistic data in machine translation	2	1	1	4
Petr Homola: Syntatic analysis in machine translation	2	2		4
Anna Nědolužko: Rozšířená textová koreference a asociační anafora	2	1	1	4
Pavel Pecina: Lexical association measures: Collocation Extraction	2	2		4
Marie Mikulová: Významová reprezentace elipsy	2		1	3
Jiří Mírovský: Searching in the Prague Dependency Treebank	1	2		3
Zdeňka Urešová: Valence sloves v Pražském závislostním korpusu	3			3
Zdeňka Urešová: Valenční slovník Pražského závislostního korpusu PDT-Vallex	3			3
Magda Ševčíková: Funkce kondicionálu z hlediska významové roviny	1		1	2
Silvie Cinková: Words that Matter: Towards a Swedish-Czech Colligational Dictionary of Basic Verbs	1			1
Total	26	10	4	**40**
Book / Month	5.2	0.8	1.3	**2**

Anja Nedoluzhko

- **projects**: GAČR “Coreference, discourse relations and information structure in a contrastive perspective” (Šárka Zikánová), KONTAKT (Šárka Zikánová), LINDAT-CLARIN, COST-Textlink
- **coreference, discourse team**
  - managing coreference annotation in PCEDT - both in English and Czech part nominal coreference annotation is added,
  - comparison of coreferential expressions in Czech and English (with M. Novak)
  - representation of Prague discourse conception to the Czech academic community (Slovo a slovesnost, terminology) - coreference part
  - preparation of coreference chapters for a collective monogtraph of our discourse team (“Text structure: the interplay of the information structure, coreference and discourse relations”)
  - plans - Anaphora resolution in Czech and its linguistic analysis (with M.Novak et al.), GAČR proposal
  - rather plans - comparison of approaches to discourse and cohesive phenomena in Czech, English and German (with colleagues in Saarbrucken)
- **morphology** (productive prefixation in Czech and Slovak with Jarka (PBML), extended to the deeper linguistic research, semantic analysis of circumfixal intensifying derivational patterns in Russian)
- Preparation of **Russian corpus annotated with coreference** (guidelines, annotation, revision, analysis)
- **teaching** (not this year) - Variabilities of languages (with Šárka and Magda)

Jiří Mírovský

Annotation of discourse phenomena (discourse relations, coreference, TFA) in PDT or elsewhere

- *with Prof. Hajičová, Šárka Zikánová, Lucie Poláková, Pavlína Jínová, Magdaléna Rysová, Kateřina Rysová, Anja Nedoluzko, and others*

Analysis by reduction performed on analytical trees

- *with Markéta Lopatková, Vladislav Kuboň*

Publication of data

- PDT 3.0, PDTSC (*Maruška Mikulová*)
- PDT 3.x?, PCEDT?

Jiří Mírovský

COST–TextLink (2014–2017)

- an interlinked database of connectives in many languages, with pointers to corpus examples
- money only for travelling (but COST–cz?)
- *with Prof. Hajičová*, *Šárka Zikánová*, *Lucie Poláková*, *Pavlína Jínová*, *Magdaléna Rysová*, *Kateřina Rysová*, *Anja Nedoluzko*

Teaching

- practical sessions for Markéta Lopatková's lectures about PDT (NPFL075)

Public Service

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

- i.e. dictionaries from Lingea, Adobe Acrobat

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

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

David Mareček

Teaching:

Selected Problems in Machine Learning (with ZŽ)

- Bayesian inference
- Gibbs sampling

Supervision:

1 bachelor student defended this year

0 left

Research:

Unsupervised grammar induction: parsing, tagging

Supervised dependency parsing

Projects:

Postdoc GAČR, QTLeap, HamleDT

PostDoc GAČR

Sentence structure induction without annotated corpora

Duration: 2014 - 2016

Budget: 1,416,000 CZK

Plan:

2014 - Grammar induction from supervised POS tags

2015 - Grammar induction without supervised POS tags (using word clustering, POS tag induction)

2016 - Employing grammar induction in applications (machine translation)

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 1.9 mil. CZK salaries (3.65 full contracts)
  - 762 th. CZK other costs
- PRVOUK (research money)
  - ca 3.0 mil. salaries (5.3 full contracts)
  - 350 th. CZK other costs
- Specific Research (?)
  - 150 th. CZK other costs
- Rozvojový program (teaching money)
  - 190 th. CZK salaries
  - 110 th. other costs

Markéta Lopatková – Teaching (1)

"Teaching projects":

- Master program in Mathematical Linguistics I3
  ("teacher responsible for the program")

- EM LCT (Language and Communication Technologies)
  together with Vladislav Kuboň
  6 students for 2014-15 (+ 1 scholarship)
  - new phase: 2013-2018

- 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 Jiří Mírovský
- Mathematical Methods in Linguistics ???

Supervising:

- 4 PhD students, 1 Master student

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, Trento, Malta, Utrecht, Groningen

Jindřich Libovický

- a PhD student, supervised by Pavel Pecina
- working on the CEMI project
  - scene text recognition – decoding string, processing the outpus of the recognition
- interested in strucutred prediction, want to learn more about deep learning
- part-timer at IBM Watson
  - dialog state tracking, anaphora and ellipsis resolution in dialogs

Veronika Kolářová

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

- Work on topics of the post-doctoral GAČR project

  "Systematic, economical and corpus-based description of valency properties of Czech deverbal nouns (theory and practice)" (P406/12/P190)
  - 2012-2014
  - principal investigator
- Main topics of interest
  - –Valency of Czech deverbal nouns
  - –Support verb constructions
- Current work
  - –Preferences in co-occurrence of participants modifying Czech nouns of communication

Veronika Kolářová

Plans for academic year 2014/2015

- Stay in England
  - Warwick (together with my husband and children)
  - Centre for Corpus Research
    - University of Birmingham
    - Study stay / research fellowship
  - We will leave on 3rd October and return in August 2015
  - So, bye for now and have a nice time at UFAL

☺

Natalia Klyueva

- PhD thesis: “Linguistic aspects of Machine Translation between Czech and Russian” (supervisor Vladisav Kuboň)
- Collecting tools and data for MT systems
- Error analysis and classification:
  - Internal: Česílko (Petr Homola), TectoMT (Martin Popel a Zdeněk Žabokrtský), Moses (with Karel Bílek)
  - External: PC Translator and Google
- Comparing SMT and RBMT
- Valency discrepancies between Czech and Russian, Czech-Russian valency dictionary

Kettnerová Václava

Grants:

- VALLEX: Delving Deeper: Lexicographic Description of Syntactic and Semantic Properties of Czech Verbs
- LINDAT/Clarin

Research interest:

valency (esp. of verbs), diatheses, alternations, lexical-semantic representation, semantic classification of verbs, light verbs

At present work on:

- Annotation of alternations in VALLEX
- Rules for the description of alternations for the grammar component of VALLEX
- Adoption of rules for diatheses for the grammar component of VALLEX
- Linking VALLEX and PDT-VALLEX, FrameNet (Eda Bejček)

Filip Jurčíček

- Main activities in the last year
  - Teaching:
    - STATISTICAL DIALOGUE SYSTEMS
      - No one got interested in course
    - BAYESIAN INFERENCE
      - Lectures - 2 hours per week
      - 2 students
    - MASTER THESES
      - Ondrej Platek
        - KALDI ASR integration into Alex
        - Placed second in SVOC of JCMF in applied informatics
  - Projects:
    - VYSTADIAL → Research into spoken dialogue systems

VYSTADIAL

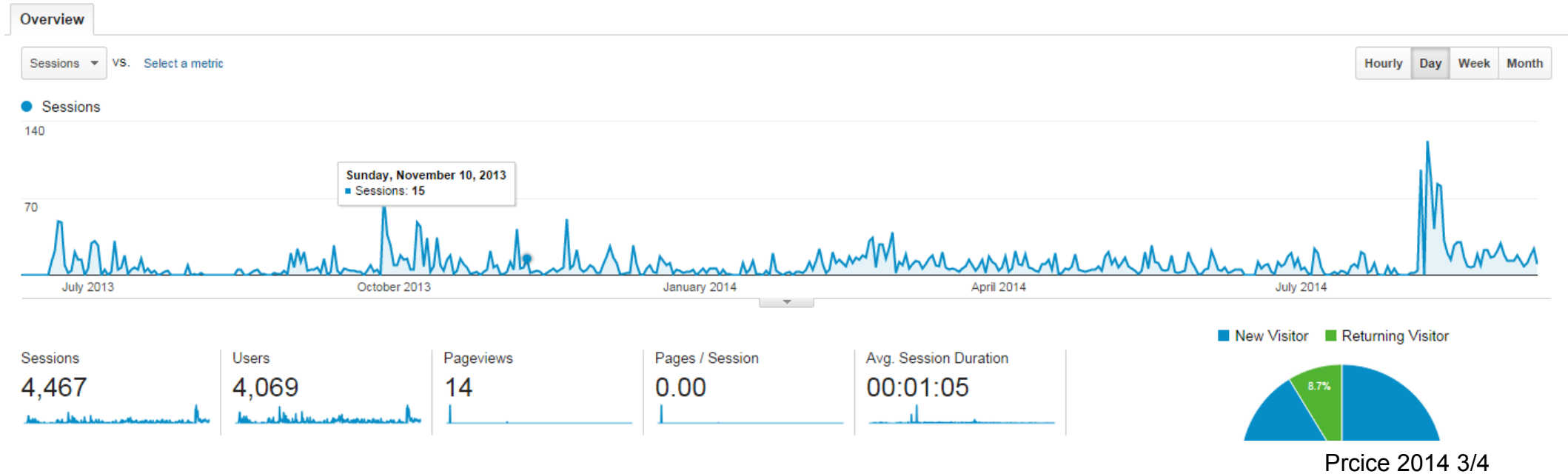
- PI: Filip Jurčíček
  - Development of statistical methods for spoken dialogue systems
  - 1.4.2012 – 31.12.2016
  - Funding for 4 PhD students

- Collaborators:
  - Ph.D. students: Ondrej Dusek, Lukas Zilka, ~~Matej Korvas~~
  - to be Ph.D. students: Ondrej Platek, Mirek Vodolan, Ahmad Agha Ebrahimian
  - Master students: Odrej Klejch, Martin Vejman, Vendula Michlíková

- Published:
  - 1x LREC, 2x SigDial, 3xTSD [RIV ;-)]
  - 2x speech datasets (English, Czech),
  - 2x software (ASR training scripts, “uzita techologie” online ASR decoder)

Main Activities

- Improvement of Public Transport Information system
  - 800 899 998 phone number
  - You can ask about
    - public transport connections in ~~Prague~~ the Czech Republic
    - weather forecast, time
  - So far we, have collected approx. 3000 calls from real-users
  - Deployed a new ASR system



SigDial 2015

- Mostly dialogue and discourse conference
- Will be held in Prague and organised by UFAL
  - 2. - 5. 9. 2015
- Volunteers are welcomed
- Local chairs:
  - Me, Prof. Eva Hajicova

Michal Josífko

- working at ÚFAL since February 2014
- technical staff for LINDAT/CLARIN project
- work in team with Pavel Straňák, Loganathan Ramasamy, Ondřej Košarko, Jozef Mišutka and Amir Kamran (ordered by geographical distance)
- maintaining and developing the repository
- maintaining the services portal
- technical help for integrating the services
- currently working on local ÚFAL installation and customization of KonText (ÚČNK) corpus manager and making ÚFAL corpora available via it

Pavlína Jínová

- since 2009 member of discourse group led by prof. Hajičová and dr. Zikánová
- annotation of discourse relations, genres, multiword discourse markers (AltLexes)
- interest: grammatical properties of discourse connectives, relationship syntax – discourse structure
- besides: since 2011 half-time teacher of Czech grammar at Faculty of arts, Charles University in Prague
- topic of Ph.D. thesis – description of Old Czech common nouns declension for automatic morphological analysis of text

Martin Holub @ UFAL

- **background** ~ computer science
- **working at UFAL** ~ 2001–2005, 2009–now
- **researcher & teacher**
- **current research interests**
  - machine learning & its applications in NLP
  - models for automatic text classification
  - lexical semantics ~ computational aspects
  - automatic procedures for lexical disambiguation
  - distributional models of lexical semantics

Teaching

- Programming for mathematicians
  ~ basic courses

- NPFL 054: Introduction to Machine Learning

  ~ focus on NLP applications
  ~ using R

- NPFL 081: Practical Foundations of Probability Theory and Statistics

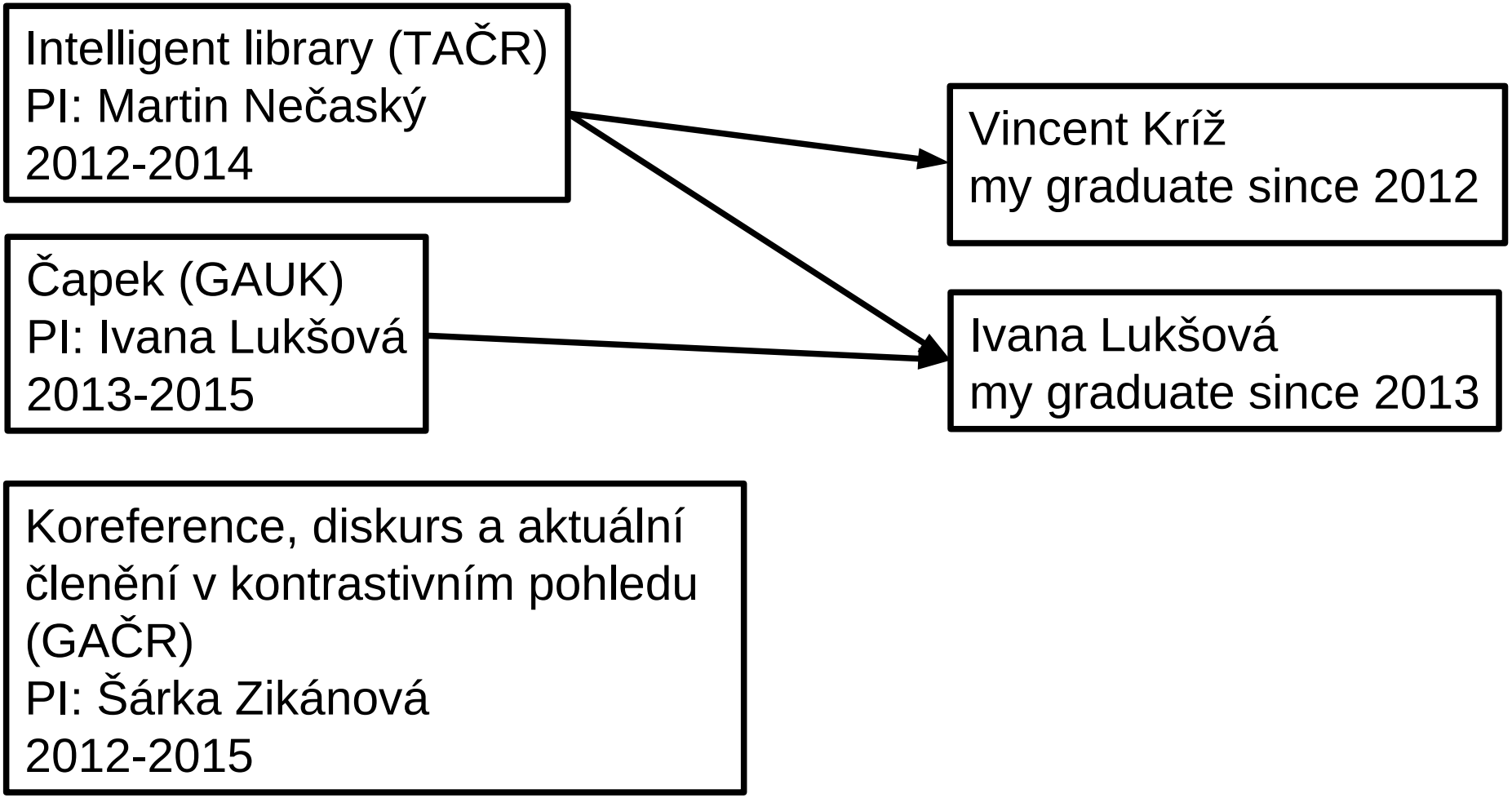
  ~ practical, supportive course
  ~ including practical introduction to R

Current research projects

- Lexical Disambiguation of English Verbs
  - Supervised models for verb context classification
  - Analysis of morpho-syntactic and semantic features
  - Heuristic algorithms for feature selection
  - Automatic verb arguments extraction
  - Selectional preferences

- Native Language Identification
  - Supervised models for text classification
  - Typical collocations and mistakes of non-native speakers

Barbora Hladká



Experience to share

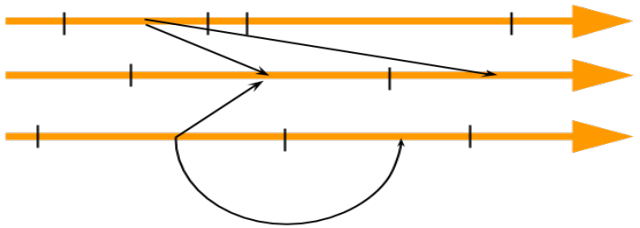
- **Certified methodology**
  - INTLIB
  - Ministry of the Enviroment of the Czech Republic
  - May 1, 2014 – initial negotiation
  - still in progress
- **Submission to Language and Linguistics Compass**
  - on-line journal indexed by Scopus
  - a paper with Martin Holub
  - submission: Nov 18, 2013; review #1: Jan 12, 2014; review #2: Jul 1, 2014; revision by Jul 30, 2014
  - still in progress

Experience to share

- **Making a contract with Datlowe, s.r.o.** – coordination
  - tokenization-ma-tagging-parsing-ner + models + PDT3.0 + CNEC2.0
  - Feb 24, 2014: a very first version of the contract
  - May 7, 2014: a final package from UFAL to Datlowe
  - Still in progress – I can see the light at the end of the tunnel, ufff.

Petra Galuščáková

- Information retrieval in audio-visual archives
  - Retrieval of the **precise beginning and end** of the relevant **segment** of the recording
    - Useful for long recordings
    - Browsing through the collection using **hyperlinks** between segments



MediaEval 2014 Search and Hyperlinking Shared Task

- **Search**: Locate a passage of the recording relevant to a given textual query in a large archive of audiovisual recordings
- **Hyperlinking**: Find more passages similar to a selected one
- 2000 hours of BBC broadcast
- Information Retrieval – using subtitles and transcripts
- Visual Similarity – in cooperation with KSI SIRET Group
- Audio Similarity – using provided prosodic features

Malach

- **Interviews** with the survivors and other witnesses of the Holocaust
- 600 hours of Czech recordings
- 5436 manually **annotated** segments
  - Each segment has manually assigned topic
- ASR **Transcripts** by ZČU
- Expected to be released during the autumn

Personal

- **2nd year Ph.D. student**
- Supervisor: Filip Jurčíček
- Thesis: **Natural language generation in dialogue systems**
- Master's at ÚFAL (2010) & Faculty of Arts (in German, 2013)

Grants (1)

VYSTADIAL (spoken dialogue systems, **Alex** framework)

- Transport + weather information – rule-based SLU, DM, NLG
- Call **800-899-998** and test it!

QTLeap (deep translation)

- Improving **TectoMT**
- Dutch analysis, English generation

Grants (2)

Khresmoi (medical translation, till August)

- **MTMonkey** – distributed translation web service
- Preparing data sets

GAUK (my thesis – natural language generation)

- Generating t-trees from dialogue acts

Other recent works

- Flect – morphological form generation
- Converting t-trees to AMR (JHU Workshop)
- Automatic valency frame assignment
- Slovak t-layer analysis (for MWE detection)
- Valency lexicon comparison (Czech & German, Faculty of Arts)

Silvie Cinková

- **Verb context analysis for lexical disambiguation**
  - patterns of verb usage, drawing on CPA
  - Rule-based collocate extraction system GRASS
  - Error analysis
    - Martin Holub, Ema Krejčová, Vincent Kríž
- **Czech TR -> AMR**
  - Rule-based, using TreeSurgeon
    - Ondřej Bojar, Ondřej Dušek, Roman Sudarikov, Zdeňka Urešová
- **SEMEVAL 2015- A CPA dictionary-entry-building task**
  - manual annotation of evaluation data
    - Vit Baisa, Jane Bradbury, Ismail El Maarouf, Patrick Hanks, Adam Kilgarriff

Ondřej Bojar

- Chiméra (Moses+TectoMT+Depfix) už podruhé pobila Google.
- Studenti: Bushra, Aleš, nově bude Roman Sudarikov
- Studentský projekt TextAn (obhajoba bude v pátek):
  - Na přání Policie ČR, podpora anotace policejních zpráv pro lepší analytiku.
  - Vloží se dokument, ručně opraví strojově označené pojmenované entity, samo to navrhne i zjednoznačnění objektů (K. Novák = Karel N. = …).
  - Vztahy mezi objekty (K.N. ukradl auto) umí TextAn pouze ručně, Zdeňkův diplomant strojově.
- Výzkum:
  - AMR, wikifikace (přiřazování jednoznačných identifikátorů zmínkám v textu).
  - Překlad vstupu hnusného a hnusnějšího:
    - Lokalizace SW (věty obsahují markup) pro IBM.
    - Twitter Crowd Translation (http://quest.ms.mff.cuni.cz:14780/)
      - Podpora pro ruční překlad tweetů (a hlavně sběr dat).
      - MT z ukrajinštiny a ruštiny, s hashtagy, překlepy.
- Služba vlasti: timesheety, maratónské PBML.

Eduard Bejček

Projects:

- VALLEX (GAČR)
  - Vallink (linking of valency lexicons), thesis
- PARSEME (COST + MŠMT)
- support with LaTeX

Research interests:

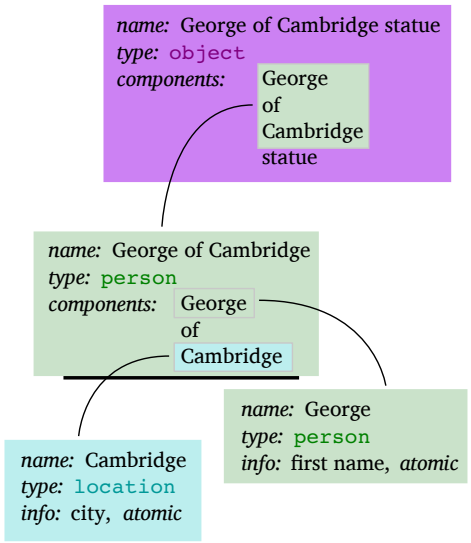
- Valency lexicons
- Multiword expressions

Eduard Bejček

PARSEME

- 30 european countries,

  29 languages
- 4 year period (2013–2017)
- meetings twice a year, short term interships



Plans together with Pavel Straňák:

- capture the inner structure of MW named entities
- MWE identification in PCEDT

Petra Barančíková

- PhD student, finishing my 2nd year, advisor Markéta Lopatková

- Thesis topic: Sentence paraphrasing for MT evaluation

- Research interests: paraphrasing, MT evaluation