

Short Presentations

ÚFAL seminar at

Malá Skála

15.-17.9.2011

Eva Hajičová

Malá Skála, září 2011

Main topics of interest

- (A) Discourse structure
- (B) Topic-Focus articulation (information structure, esp. its semantic relevance)
- (C) Dependency structure of sentences
- => and the relation of (A), (B) and (C)
- Czech-English contrastive studies

Grants:

- Výzkumný záměr (ending by Dec. 2011)
- LINDAT-Clarin
- GAČR (Principal Investigator): discourse (ending by Dec. 2011)
- International: Czech-US (Univ. of Pennsylvania) scientific cooperation (Principal Investigator) – Discourse annotation – till Dec. 2012

Plans (if I can still speak about plans)

- To use the PDT annotation of syntactic relations, TFA, and coreference relations to study the interaction of these relations, esp. with regard to the development of discourse in terms of the hierarchy of salience of items of the stock of knowledge the speaker assumes to share with the addressee (dynamics of discourse)
- To look more closely on the PDT-style annotation of English, with view on TFA annotation of English

Collaborators

- Different aspects of discourse:
 - Pavlína Jínová, *Jiří Mírovský*, Lucka Poláková-Mladová, *Anja Nedoluzhko*, Kateřina Rysová, Magda Rysová, *Bára Vidová Hladká*, *Šárka Zikánová*
- Other topics (PhD):
 - Zdena Urešová
 - Kateřina Veselovská
 - Barbora Štěpánková

- Personal remark:
- My continuous offer to read your abstracts, papers etc. is NOT caused by a desire „to put my nose into your business“ or by any „imperialistic“ ambitions, but by a mere (friendly) intention to offer „my hand“ (perhaps based on a long time of experience) – so PLEASE do not fear that you bother me by asking (though please leave me more than one night for reading 10 pages)

Jarmila Panevová

- GA ČR P406/2010/0875 „Computational Linguistics: Explicit description of language and annotated data focused on Czech“ (2010–2013)
 - revision of the PDT tectogrammatical annotation
 - experimental annotation of newly introduced grammatememes:
 - pair/group meaning with selected nouns
 - resultative 2
 - resultative 1 vs. passive constructions
 - ...
 - to be implemented in PDT 3.0 (2.5?)
 - annotation by two annotators in parallel

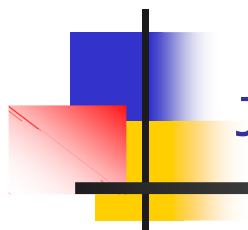
Grammatical diatheses of Czech verbs

- basic opposition of active and passive voice
 - *he cooked an excellent dinner vs. the dinner is cooked*
- further types distinguished:
 - resultative constructions
 - objective forms (grammateme value *resultative1*)
 - *je zajištěno, že přijde* `(it) is arranged that (he) comes`
 - ambiguous with passive (grammateme value *passive*)
 - possessive forms (grammateme value *resultative2*)
 - *dohodu máme podepsánu* `(we) have the agreement signed`
 - recipient diatheses
 - result of the syntactic process constituting a recipient paradigm
 - the Addressee shifted to the position of surface subject
 - *za posudek dostal zaplaceno* `(he) got payed for the review`

Publications and presentations

- on the pair/group meaning:
 - conference paper (with M.Ševčíková and Z. Žabokrtský) at the International Workshop on Treebanks and Linguistic Theories (TLT 2010), December 2010, Tartu, Estonia (in proceedings)
 - oral presentation at *ÚFAL Seminar of Formal Linguistics*, March 7, 2011
 - journal article (with M.Ševčíková) in *Slovo a slovesnost*, 72, 2011, pp. 163–176
 - journal article (with M.Ševčíková and L. Smejkalová) in *Prague Bulletin of Mathematical Linguistics*, 96, 2011, pp. 27–48
- conference paper on grammatical diatheses and other issues at the Conference on Dependency Linguistics (Depling 2011), September 5–7, 2011, Barcelona, Spain (in proceedings) – with M. Ševčíková
- articles in festschrifts
 - Festschrift for Prof. J. D. Apresjan, Moskva
 - Festschrift for Prof. M. Grochowski, Warszawa (with M. Mikulová)
- 5 book chapters on grammatical issues in the *Chapters from the Czech Grammar* (F. Štícha, ed.)

- oral presentations
 - on PDT annotation – for PhD students of linguistics at the Faculty of Philosophy and Arts, Charles University in Prague, October 21, 2010 (with M. Ševčíková)
 - on the revision of grammatemes for PDT 3.0 – at the *ÚFAL Seminar of Formal Linguistics*, November 1, 2010 (with M. Ševčíková)
 - on corpora of Czech – for the Kruh přátel českého jazyka, March 9, 2011 (with M. Ševčíková)
 - on Czech syntactical studies in the last half of the 20th century from the point of view of Prague and Brno at the conference “100 years of the Institute for Czech Language”, June 2, 2011 (with P. Karlík)
 - at the 6th Annual Meeting of the Slavic Linguistics Society in Aix-en-Provence, September 2, 2011 (with M. Mikulová)



Jarmila Panevová

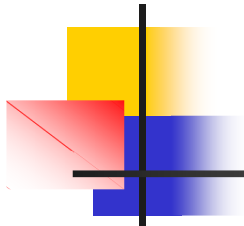
Malá Skála, 15.-17. 9. 2011

Mluvnice současné češtiny 2

Syntax na základě anotovaného korpusu

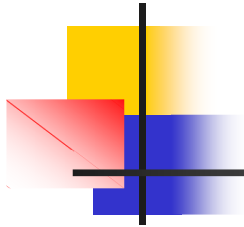
(Grammar of the Contemporary Czech 2

Syntax on the basis of the annotated corpus)



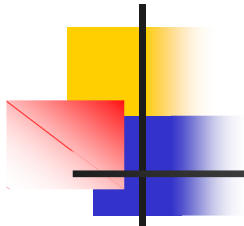
Malá Skála, 15 - 17.9.2011

- Volume 1 – **Václav Cvrček et al., How we write and speak**, Karolinum: Praha, 2010
- Volume 2 – J. Panevová et al. is expected since 2009
- (al. means: E. Hajičová, M. Lopatková, J. Štěpánek, M. Mikulová, M. Ševčíková, V. Kettnerová and any volunteer actively participating)



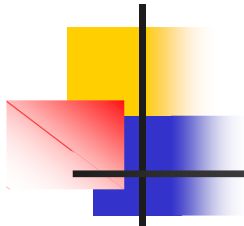
Malá Skála, 15.-17.9.2011

- **Content** (1st version May 2006, modified version June 2009, revised version – last weekend)
- Ready:
- 0. Introduction *2 p.*
- 1. Dependency approach to Syntax *15 p.*
- 2. The types of Dependency Relations *29 p.*
- (not complete, Semantic Diathesis are supposed to be included – M. Lopatková, V. Kettnerová)



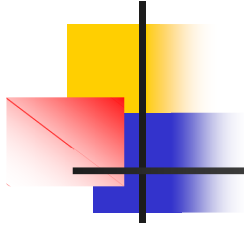
Malá Skála, 15.-17.9.2011

- 7. Information Structure of Sentence and Word Order (E. Hajičová, K. Veselá) *cca 10 p.*
- In progress:
- 5. Grammatical coreference (control, quasi-control, reflexivity, relative clauses)
(papers of J. Panevová, V. Kolářová)
- 6. Ellipsis (delations) and similars
(thesis of M. Mikulová, papers J. Panevová, M. Mikulová)



Malá Skála, 15.-17.9.2011

- 7. Meanings of the morphological categories and their participation on the sentence structure
- (grammatical diatheis – J. Panevová, *19 pages*,
- semantical parts of speech („sempos) – M. Ševčíková, *10 pages*,
- noun number, degrees of adjectives and adverbs,
- sentential and verbal modality in simple and complex sentences
- ?reciprocity as a syntactical diathesis (*shift on another place?*))
- *Papers of J. Panevová, M. Ševčíková, M. Mikulová*



Malá Skála, 15.-17.9.2011

- Completely missing:
- 4. Coordination, apposition, parenthesis
- Help of any type welcome (tree pictures, registers, bibliography items etc.)!

Eduard Bejček

Lexemann

- MWEs annotated in PDT
- manual data cleansing of annotations for **PDT 2.5**

Vallink

- link corresponding frames from PDT-VALLEX and VALLEX
- spin-off: linking old and new VALLEX (1.0 and 2.5)
 - resulted in **VALLEX 2.6**
 - a few corrections
 - VALEVAL example sentences for all annotated verbs
 - VALEVAL annotation in a format ready for machine learning [will be in a few days]
 - <http://ufal.mff.cuni.cz/vallex/2.6/> [will be in a few days]

Vallex 2.6

označovat^{impf}, označit^{pf}

1 ≈ **impf**: dělat značku **pf**: udělat značku

-frame: **ACT**₁^{obl} **PAT**₄^{obl} **MEANS**₇^{typ} **BEN**₃^{typ}

-example: **impf**: označovat jim / si řádek křížkem **pf**: označil Petrovi tašku nálepkou

-usage in ČNK: **impf**: Nálepky, kterými musejí řidiči označovat vozy jezdící po dálnicích a silnicích dálničního ... (more...)
pf: ... na níž si lidé mohou nechat označit pohlednice a dopisy příležitostným razítkem k ... (more...)

-rfl: pass: **impf**: výsledky se označují dvojím podtržením **pf**: křížkem se označily chyby

-rcp: ACT-PAT: **impf**: označovali se navzájem pro sebe **pf**: označili sebe navzájem pro sebe

2 ≈ **impf**: nazývat; určovat **pf**: nazvat; určit

-frame: **ACT**₁^{obl} **PAT**₄^{obl} **EFF**_{jako+4,za+4,jako+adj-4,za+adj-4}^{obl} **BEN**₃^{typ}

-example: **impf**: označoval někoho jako blbce / za blbce **pf**: označit jim někoho za vraha; označoval někoho jako blbce / za blbce

-usage in ČNK: **impf**: Za šílence označují přátelé Martina Nicholase, anglického pěstitele a ... (more...)
pf: Dá se označit jako psychedelická komedie. (more...)

-rfl: pass: **impf**: každý, kdo se chová jako on, se označuje za blbce **pf**: protivník se označí za zrádce

-rcp: ACT-PAT: **impf**: označovali se navzájem za viníky **pf**: označili se navzájem za škůdce společnosti

-class: appoint verb

3 ≈ jen **označovat**^{impf}
znamenat

-frame: **ACT**₁^{obl} **PAT**₄^{obl}

-example: tento termín označuje stabilitu

-usage in ČNK: **impf**: ... šumí, nýbrž slovo šuma označovalo neprůchodný les. (more...)

Valeval

označovat

1

Dálniční nálepky na rok 1996 jsou zelené PRAHA (ČTK) - Zelené s červeně vytištěným rokem 1996 jsou dálniční nálepky pro motorová vozidla na příští rok, které se dnes prodávají již pátým dnem. Nemění se cena nálepek, která zůstává 400 korun pro osobní vozidla, 1000 korun pro vozidla do hmotnosti 3.5 tuny a 2000 korun pro vozidla s vyšší hmotností.

#46 **Nálepky, kterými musejí řidiči označovat: 1 vozy jezdící po dálnicích a silnicích dálničního typu, lze obdržet u stálých stanovišť označených nálepkou "Prodejní místo" se symbolem dálniční známky pro rok 1996.** (S/NWS/1995/hnh5:26738-p2s3)

Alfred se stal králem Wessexu, jihozápadní oblasti anglosaské Británie, r.871 a roku 886 se pak zmocnil Londýna a stal se králem celé Anglie. Přesný účel předmětu zůstává nejasný. Možná šlo o ústřední šperk na korunu, anebo se nosil jako brož či přívěsek.

#56 **Mohlo jít dokonce i o špičku ukazovátka "aestel", jímž se označovaly: 1 řádky v rukopise.** (S/B/1996/poklady:003-p70s9)

[loga DOLBY STEREO a DOLBY STEREO SPECTRAL RECORDING] Obdobně jako kina jsou označovány také filmové programy s ochrannou známkou DOLBY STEREO a doplňkovým nápisem "ve vybraných kinech". Tento symbol se objevuje vždy na začátku či na konci filmového programu.

#59 **Stejně tak jsou označovány: 1 filmové plakáty a propagační materiály.** (S/J/1997/film0597:016-p16s3)

2

Odpověď z 31. 8.: Zpěvák skupiny Verve se jmenuje Richard Ashcroft. Výherci: Josef David z Prahy, Josef Trýnoha z Časlavi a Kateřina Tothová z Mokřin. Nebezpečný koníček

#2 **Za šílence označují: 2 přátelé Martina Nicholase, anglického pěstitele a milovníka tarantulí.** (S/NWS/1998/pr980921:154-p2s1)

Aero Vodochody dodává prostřednictvím Eagle Group letouny L-39 (jsou vybaveny elektronikou izraelské zbrojovky Elbit) do Thajska. Elbit je sice soukromou firmou, nicméně na řízení izraelských zbrojovek se podílí stát. Podle informací LN předcházeli pátečnímu ustavení komise expertů, která připraví vládě podklady k výběru partnera pro L-159, dopis premiéra V. Klause ministru obrany A. Baudyšovi i šéfu Aera Z. Chalupníkovi, aby zohlednili ekonomické aspekty projektu.

#4 **Ten odborníci označují: 2 za ojedinělý.** (S/NWS/1994/ln94209:017-p3s5)

6.1. Nehnisavá onemocnění K nehnisavým infekcím kůže patří především onemocnění, která charakterizuje generalizovaný výsev kožních eflorescencí označovaný jako exantém .

#5 **U některých infekcí je exantém pravidelným a nápadným příznakem nemoci a tyto nákazy označujeme: 2 jako exantémové .** (S/B/1997/havlik:001-p1375s2)

Krátce Ostrava: Severomoravská organizace ODA bude prosazovat svého šéfa Miroslava Tošera na místo jednoho z místopředsedů strany. Na včerejším jednání organizace to potvrdil její manažer Jan Večerek.

#6 **Starosta Kravaň Tošer bývá označován: 2 za jednoho z nejvýraznějších kritiků dosavadního vedení ODA.** (S/NWS/1997/ln97016p:005-p3s3)

PRAHA - Návrat Viktora Černomyrdina do premiérského křesla podle analytiků nevyřeší ruskou finanční krizi, která se vyhrtila v polovině minulého týdne. Důležité je, aby vláda rychle pokračovala v reformách, shodují se všichni. Ruská krize se totiž projevuje i v okolních zemích.

#8 **Kvůli ní oslabil minulý týden polský zlotý o osm procent, padá i maďarský forint a měny ostatních zemí, které investoři**

Ondřej Bojar

- Research
 - Moses:
 - mainly tuning for English-Czech
 - also "beta-testing", occasional minor extensions, organizational things (moving repo to Git)
 - Data:
 - co-supervising CzEng and other corpora
 - MT evaluation:
 - finding flaws in both automatic and manual methods (see our WMT paper, ask Martin Popel)
 - blue skies stuff (i.e. long term interest):
 - hybrid deep and phrase-based search spaces; MT into morphologically rich languages in general

Ondřej Bojar, cont.

- Research management:
 - supervising our part in EuroMatrixPlus
 - co-writing grant proposals for various agencies (EU FP7: MosesCore in negotiation; two TACR grants)
- Teaching:
 - a seminar on MT
 - supervising mainly Bc. and Mgr. theses on various topics

Silvie Cinková

- **Prague English Dependency Treebank**
 - manual tectogrammatical annotation, checking, release documentation
 - *with Eva Fučíková, Josef Toman & many annotators*
- **Vector space model of semantic preferences in English verbs, exploiting Corpus Pattern Analysis**
 - human lexicography, P. Hanks
 - clustering of concordances according to similarity
 - manual annotation + lexical description of each verb in experiment, interannotator disagreement analysis \Rightarrow data preparation for ML experiments
 - *with Martin Holub, Lenka Smejkalová, Anna Lauschkmannová and Vincent Kríž*

ÚFAL Projects

FAUST – Czech analysis and synthesis to improve MT output

- ▶ MT output often contains correct senses with wrong grammar
- ▶ Goal: Improve the grammar by analyzing and generating back
- ▶ **Czech NLG**: more statistical, more language-independent, dealing with incomplete information

CzEng 1.0

- ▶ improved formemes set, automatical functor assignment

German-Czech reference translations of WMT '11 test set

- ▶ organizing and checking human translations

Other Topics of Interest

Including:

- ▶ machine learning approaches to NLP, combination with rules
- ▶ valency theory
- ▶ corpus linguistics
- ▶ German philology, especially linguistics
- ▶ Dutch language



Nathan Green
green@ufal.mff.cuni.cz





Background

Activity	Organization	Dates
Undergraduate	North Carolina State University	1999 - 2003
Internship	IBM	2000
Masters	George Washington University	2003 - 2006
Software Engineer	Bureau of Labor Statistics	2003 - 2006
Fulbright Fellowship	University of Iceland	2006 - 2007
Graduate Research	North Carolina State University	2008 - 2010
PhD	Charles University/ CLARA	2010 - ...





Research 2010-2011

● Dependency Parsing

- Nathan Green, Effects of Noun Phrase Bracketing in Dependency Parsing and Machine Translation, Association for Computational Linguistics (ACL) Student Session, Portland, Oregon, 2011
- Martin Popel, David Marecek, Nathan Green and Zdenek Zabokrtsky, Influence of Parser Choice on Dependency-Based MT, Proceedings of the Sixth Workshop on Statistical Machine Translation, Edinburgh, Scotland, pp. 433-439, 2011

● Collaborative Environments

- N. Green, P. Breimyer, V. Kumar, and N.F. Samatova, PackPlay: Mining semantic data in collaborative games, Association for Computational Linguistics (ACL) in the Proceedings of the Fourth Linguistic Annotation Workshop (LAWIV), Uppsala, Sweden, pp. 227-234, 2010

● Data Mining

- Practical Graph Mining with R (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series)

- Chapter 4: Link Analysis

- Chapter 5: Graph-based Proximity Measures





- Clara Training Activities

- Prague - Consolidating and Harmonizing Treebank Annotation
- Budapest - Processing Morphologically Rich Languages
- Croatia - Industrial Career Training Course on Product Planning for Next Generation Information Access Technology Solutions

- Conferences

- ACL 2011- Portland
- EMNLP/WMT 2011 Edinburgh
- MT Marathon - Trento





Future Research

- Continue looking at dependency parsing and its influence on statistical machine translation
- Start researching new techniques for unsupervised dependency parsing



Martin Holub

- **full time researcher & teacher**
- **background:**
 - **computer science**
 - **PhD in software engineering**
 - **at UFAL: 2001-2005, 2009-now**
- **current research interests:**
 - **lexical semantics (computational aspects)**
 - **distributional models of lexical semantics**

Martin Holub

- **teaching**

- **programming - basic courses**
- **introduction to machine learning (in R)**
- **seminar in computational aspects of lexical semantics**

- **current research project(s)**

- **semantic annotation of English verbs**
- **selectional preferences of verb arguments**
- **distributional models of lexical semantics**
- **semantic similarity, semantic clustering**

Angelina Ivanova. Profile

Education

- **St. Petersburg State University**
Applied Informatics in Arts and Humanities
- **Free University of Bolzano-Bozen**
Charles University in Prague
Language and Communication Technologies

Angelina Ivanova. Profile

Master's Thesis

Creating a Bilingual Dictionary Using Wikipedia

- ***Development***
- ***Evaluation***
- ***Application to Machine Translation***

Angelina Ivanova. Profile

Research Interests

- *Textual Entailment*
- *Statistical Parsing*
- *User-Generated Content (e.g. Wikipedia)*
- *Semantic Web*

Pavlína Jínová

Study

- since 2009 Ph.D. study of Czech language at Faculty of Arts
- thesis: Morphological Processor and Lemmatiser of Old Czech Substantives

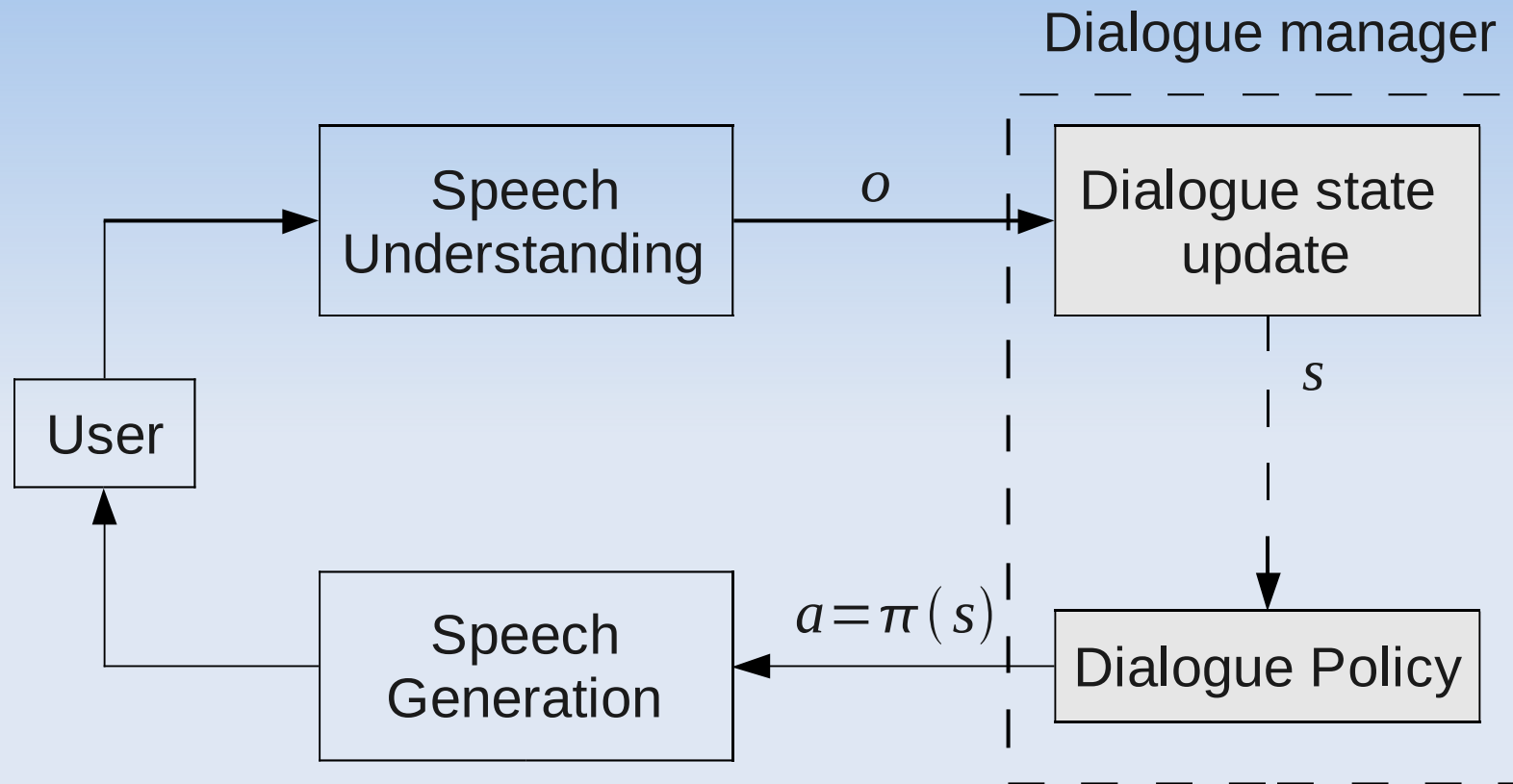
Work

- since 2009 discourse annotation project at UFAL (with Lucie Mladová, Šárka Zikánová, Magda Rysová and others)
- annotation, tuning annotation scenario, preliminary research...
- since 2010 teaching at Faculty of Arts (introductory language course, morphology)

Filip Jurcicek

- Post-Doc at UFAL
 - since this September
- Last three years have worked on
 - semantic parsing for Spoken Dialogue Systems
 - parameter estimation in statistical Spoken Dialogue systems
- Future
 - Work on META-NET

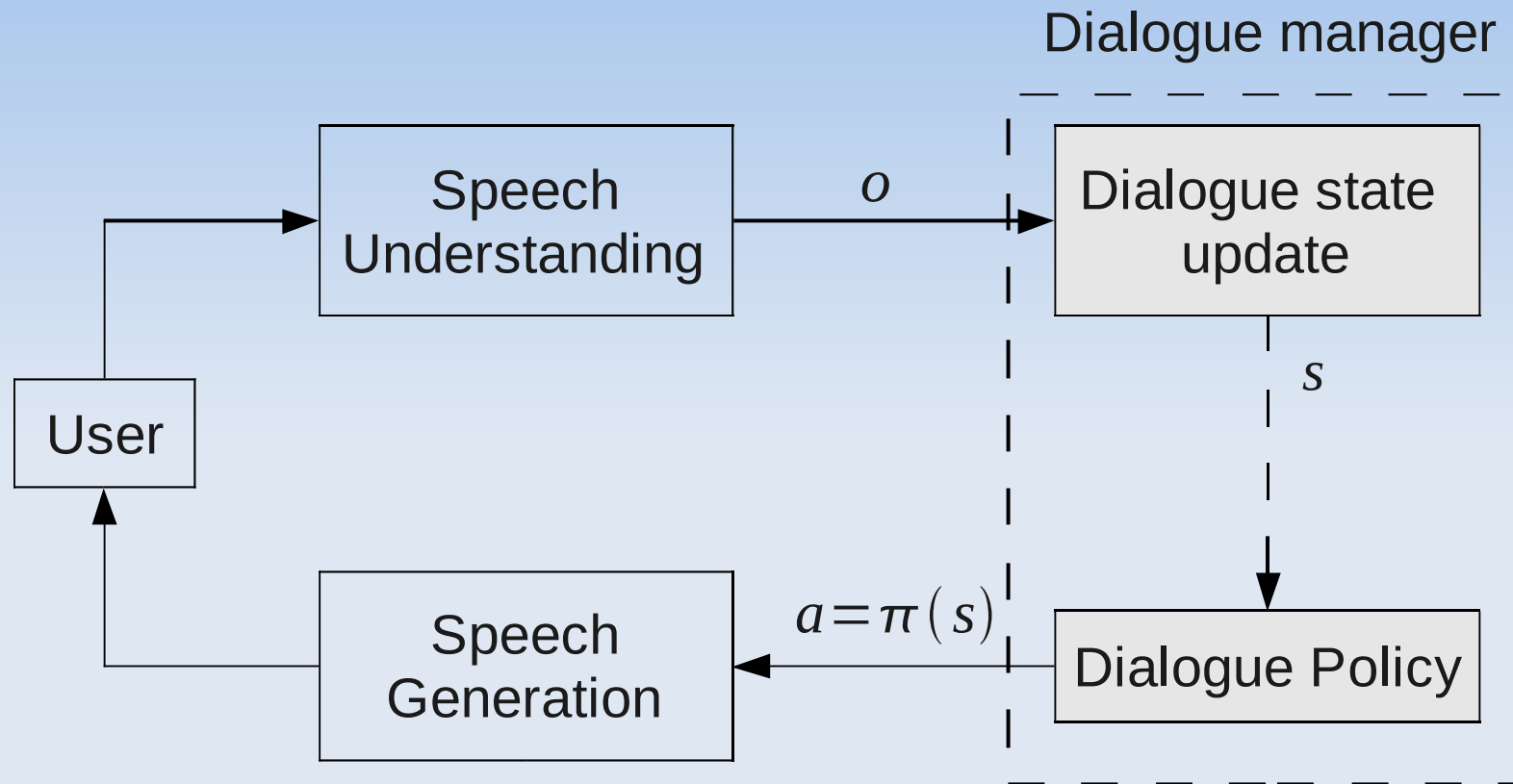
Typical Spoken Dialogue Systems



- DM updates the its dialogue state given the last user's input and the last system's output
- Based on policy, DM generates the system's output given the current dialogue state

Typical Spoken Dialogue Systems

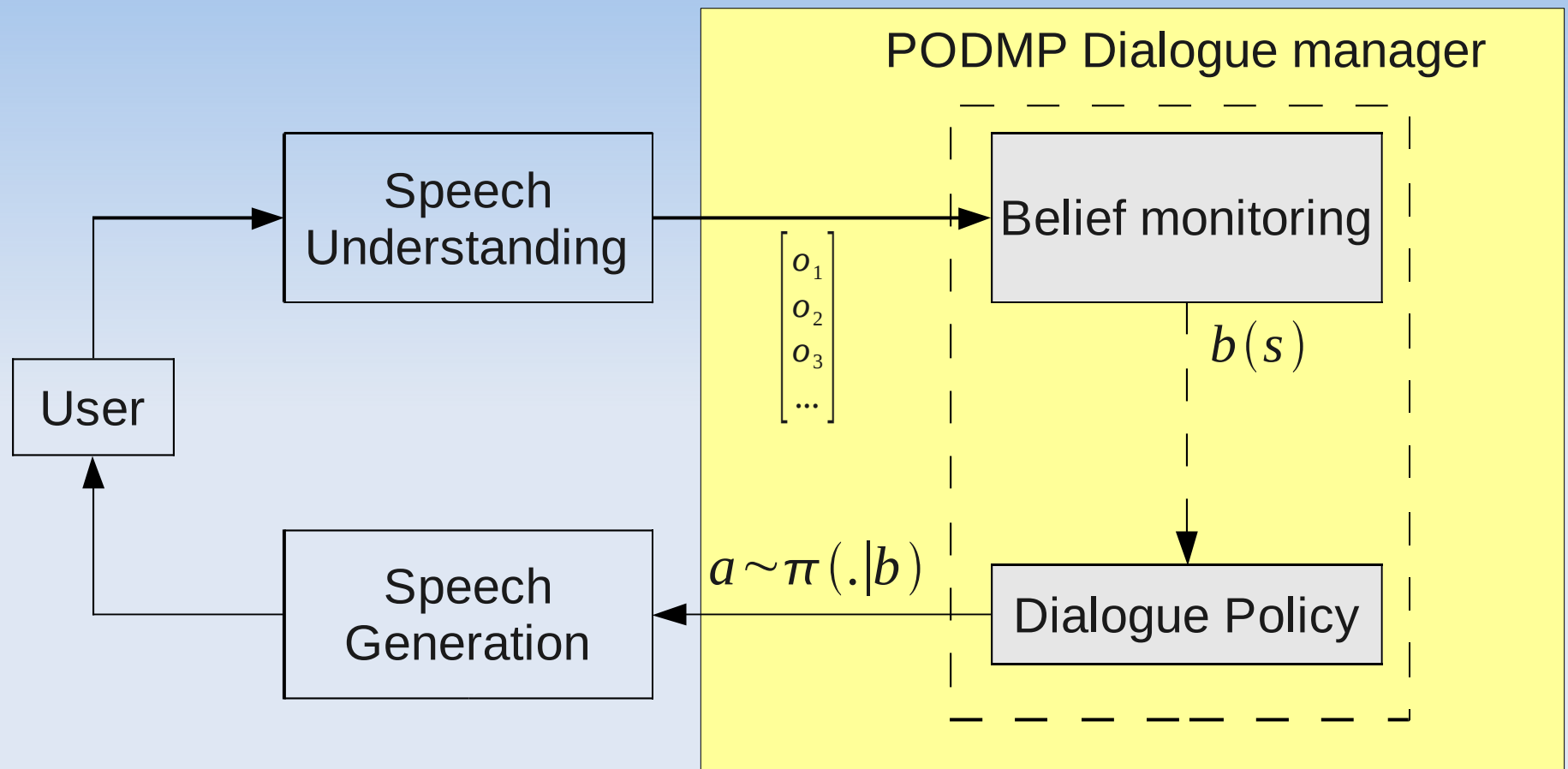
Dialogue state and policy



- Dialogue state is composed of **variables** needed to track the the porogress of the dialogue
- Policy is implemented as a sequence of **if/then decision**

Statistical Spoken Dialogue Systems

Partially Observable Markov Decision Process



- Instead of tracking the state s , the dialogue manager maintains a distribution over all states: $b(s)$
- Policy explicitly takes into the account in the uncertainty in $b(s)$

META-NET

- Speech to speech translation
 - Dialogue manager for multilingual multi-party dialogues
 - Integration of semantics and pragmatics
 - Benefits:
 - better handling of the ASR and SLU errors
 - it can bridge cultural differences
- Can actively managed dialogue based translation model improve user satisfaction?

Kettnerová Václava

- Grants: Center for computational linguistics II and Computational Linguistics: Explicit description of language and annotated data focused on Czech
- Research interest: valency (esp. of verbs), diatheses, alternations, lexical-semantic representation, semantic classification of verbs
- Linking VALLEX and PDT-VALLEX (Eda Bejček)
- Thesis: Lexical-semantic conversion in Czech (the similar relations btw. syntactic structures e.g. *The farmes loaded hay onto the truck* – *The farmers loaded the truck with hay* and their representation in the valency lexicon VALLEX)

Veronika Kolářová

- Part-time job (30%) for GAČR
 - Panevová; Computational Linguistics: Explicit description of language and annotated data focused on Czech
- Issues of valency of Czech deverbal nouns, esp. those derived from verbs with a participant expressed by prepositionless Genitive
 - adnominal counterparts of adverbial prepositionless Genitive
 - nominal constructions with two participants expressed by prepositionless Genitive

Tomáš Kraut - profile

Tomáš Kraut

Malá Skála

15th September 2011

1 My work at ÚFAL

2 My studies at ÚFAL

Tomáš Kraut as employee

- Since January 2011
- Maintaining Treex NLP Framework
 - Core of framework is already on CPAN
 - Plan to release English demo in September

Tomáš Kraut as student

- Since October 2010
- Searching for topic of thesis

Statistical modeling of lexical meaning

Future dissertation

Ema Krejčová

September 2011

My background

- ▶ study at MFF UK
Bc. - General mathematics
Mgr. - Mathematical methods of information security
Diploma thesis: *Algebraic proof systems* (2009)
- ▶ attempt at FF UK
General linguistics
English for intercultural communication

→ computational linguistics

Topic: Statistical modeling of lexical meaning

- ▶ **lexical meaning** - meaning of a word as a lexical unit in a particular context
- ▶ **modeling of meaning** - a word's meaning characterized as a set of contexts in which the word can appear
- ▶ **statistical modeling** - analyses of large corpora using statistical methods

Distributional hypothesis:

"A word is characterized by the company it keeps" (J.R. Firth)

- ▶ words that occur in the same contexts tend to have similar meanings

What is it good for?

For automatic processing and evaluation of semantic properties of lexical units:

- ▶ semantic similarity
- ▶ substitutability
- ▶ semantic relations between units
- ▶ polysemy, ambiguity → guessing the word's meaning in a given context

The aim of the dissertation

- ▶ design and implementation of a semantic computational model for the analysis and representation of lexical meaning and semantic relations between lexical units
- ▶ using statistical analyses of text corpora, algebraic methods and machine learning
- ▶ applicable for various tasks in computational linguistics

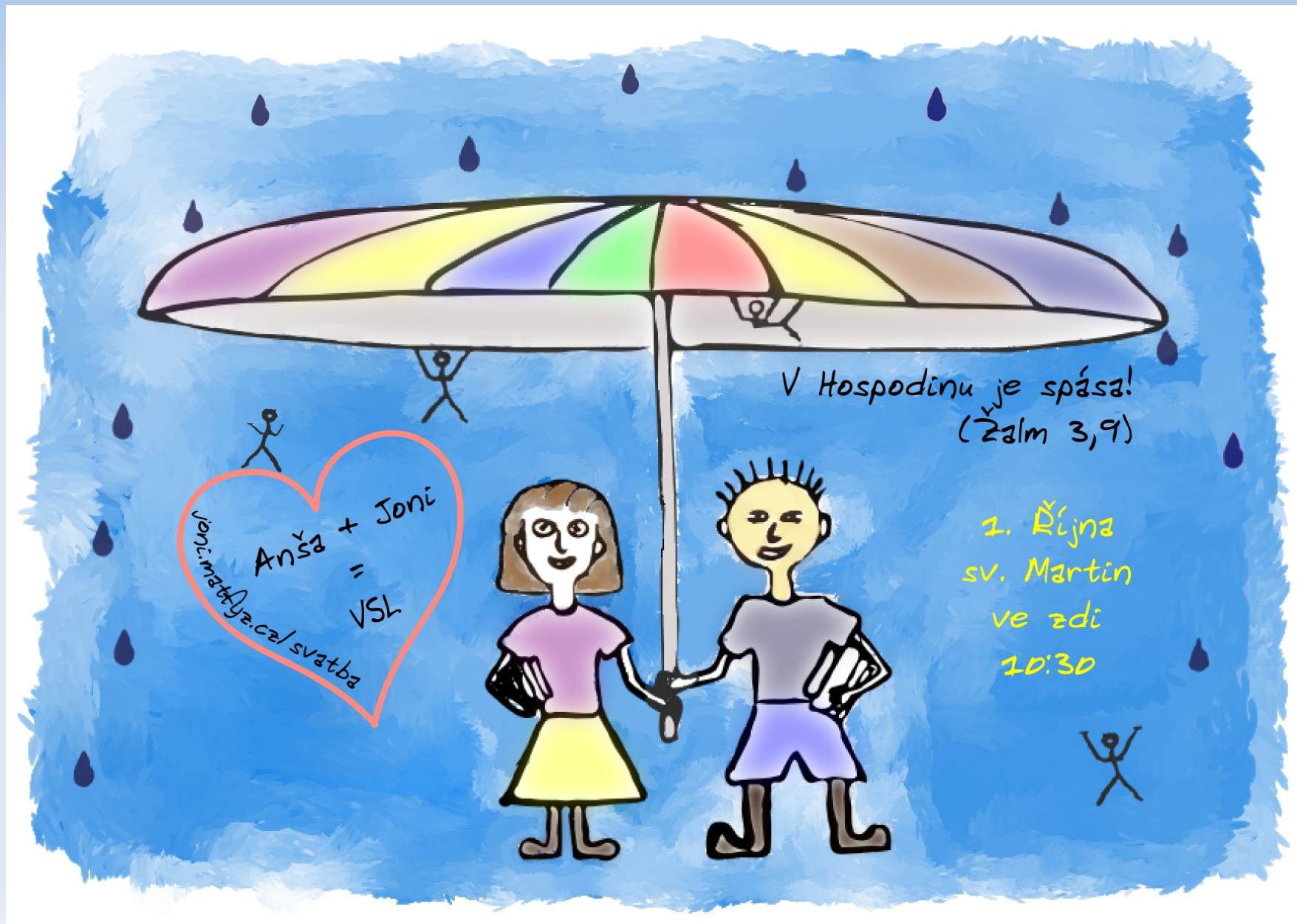
Vlád'a Kuboň

- Prrojekty
 - podílím se na vedení grantu Res Informatica,
 - spolupráce s firmami Hieronymus, Virtus
- Výzkumná činnost
 - Segmentační analýza
 - rozklad souvětí na lingvisticky motivované úseky a následná snaha odhadnout vzáj. vztahy klauzí
 - Překlad mezi příbuznými jazyky (Česílko)
 - Programový výbor několika workshopů (EAMT, IIS atd.)
- Výuka
 - Přednáška „Úvod do poč. lingvistiky ...“ česky i anglicky; dále 2 semináře; v LS organizace přednášky Aplikace NLP, cvičení Automaty a gramatiky
 - Společně s Markétou Lopatkovou řídíme mezinárodní magisterský program LCT v rámci programu Erasmus Mundus,
 - Tajemník ÚFALu pro výuku
 - Koordinuji výměnu Erasmus se Saarbrueckenem, s Koperem a s Tuebingen
 - Vedl jsem 1 diplomku (obhájena v lednu), 4 postgraduanty (od října jich bude 5), dále 1 bakalářku a 1 ročníkový projekt

Anna Lauschmannová

- Valency dictionary of Czech nouns
 - PhD dissertation, advisor Markéta Lopatková
- Corpus Pattern Analysis
 - Annotator
- Transdisciplinarity
 - project of PhD students

Anna Vernerová



Loganathan Ramasamy

Charles University in Prague

Sep 15, 2011

Outline

- 1 Research
- 2 Other activities
- 3 Publications

Treebanking

- Treebanking

- Developed a small PDT style treebank for Tamil.
- Annotation layers: (i) m-layer and (ii) a-layer.
- Data size: 600 sentences from news corpus.
- Data and documentation available at:
<http://ufal.mff.cuni.cz/~ramasamy/tamilTB/0.1/>

- Treebank normalization

- Collaborative work with others.
- Experimenting different annotation decisions in treebanks.

Other activities

- CLARA

- Course on Treebank Annotation, Dec 13-16, 2010, Prague.
- Training Course on Processing Morphologically Rich Languages, Apr 11-15, 2011, Budapest.
 - Worked on different FST tools.
 - Developed a small prototype Tamil morph analyzer based on FOMA.

- MT Marathon 2011, Trento

- Worked on integrating Translation Memory(TM) into Moses.
- Still working on it.

Publications & Talks

- ① Loganathan Ramasamy and Zdenek Zabokrtsky, **Tamil dependency parsing: results using rule based and corpus based approaches**, CICLing 2011.
- ② Loganathan Ramasamy, **TamilTB: An Effort Towards Building a Dependency Treebank for Tamil**, WDS 2011 (*to appear*)
- ③ **TamilTB: An Effort Towards Building a Dependency Treebank for Tamil**, *Talk in UFAL's Monday Seminar*, May 16, 2011.

Markéta Lopatková

Research interests / research projects:

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

Object of Research (*Výzkumný záměr*) MSM0021620838

GAČR (new application): *Delving Deeper: Lexicographic Description of Syntactic and Semantic Properties of Czech Verbs*

- Modeling of stratificational dependency-based syntax
based on 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

Markéta Lopatková (2)

"Teaching projects":

- EM LCT (Language and Communication Technologies)
together with Vladislav Kuboň
4 students for 2011-12
- CLARA (Common Language Resources and their
Applications) - a Marie Curie Action, 2009-13
Winter School (Feb. 2012)
"New developments in computational linguistics"

Plans for winter term:

University of Edinburgh, prof. Bonnie Webber
translation of pronouns for SMT (Liane Guillou)

Markéta Lopatková (3)

Teaching:

- Mathematical analysis:
~~winter~~+ summer term, 2 practical courses, BSc.
- Prague Dependency Treebank
with Honza Štěpánek

Supervising:

- 4 PhD students, 2 Master students

Others:

- Grant Agency of Charles University
committee for computer science (*oborová rada*)
- editorial board: Slovo a slovesnost, Korpus – Gramatika – Axiologie
- coordinator of Erasmus exchange – Bolzano, Malta, Utrecht, Groningen

Jiří Mírovský

- Project of annotation of extended textual coreference and bridging anaphora

Anja Nedoluzhko

annotation tool (an extension of TrEd), data management, measurement of the inter-annotator agreement, data corrections

- Project of annotation of discourse

Prof. Hajičová, Šárka Zikánová, Lucie Mladová, Pavlína Jínová

annotation tool (an extension of TrEd), data management, measurement of the inter-annotator agreement, data corrections

- Some new grammaticemes for the new PDT

Prof. Panevová, Magda Ševčíková

searches and scripts

Jiří Mírovský

- PlayCoref

Bára Hladká, Jan Kohout
data manipulation

- Public Service

- purchase and management of software (except SW from Microsoft)
- purchase and management of data published by LDC
- assistant management of Amoeba (database of employees at ÚFAL)

Jiří Mírovský

- PlayCoref

Bára Hladká, Jan Kohout
data manipulation

- Public Service

- purchase and management of software (except SW from Microsoft)
- purchase and management of data published by LDC
- assistant management of Amoeba (database of employees at ÚFAL)

- Team Building

Habanero degustation

- **2nd hottest chilli**
- **habanero and milk provided**
(bring your own cup)
- **only active participants**



Jiří Mírovský

Habanero degustation

- 2nd hottest chilli
- habanero and milk provided
(bring your own cup)
- only active participants



Scoville Scale; SHU = Scoville Heat Unit (Wilbur Scoville, 1912)

SHU	product
15,000,000	Pure capsaicin
500,000 – 5,300,000	Law enforcement pepper spray, irritant ammunition
1,000,000	Naga Jolokia (a recently discovered (~cultivated) pepper)
100,000 – 350,000	Habanero (Red Savina H. up to 580 th. SHU)
2,500 – 8,000	Tabasco sauce
0	Sweet pepper

Lucie Poláková (Mladová)

Current project:

Annotation of discourse structure on PDT (Czech texts, based on tectogrammatcs)

- Together with: Eva Hajičová, Šárka Zikánová, Jirka Mírovský, Pavlína Jínová
- Cooperation with UPenn and prof. Aravind Joshi's team (KONTAKT project)
- Annotators: Pavlína Jínová, Jana Zdeňková, Veronika Pavlíková, Magda Rysová
- Till August 2010 – 43,4% of the treebank annotated (for explicit discourse connectives and their relations)
- Till August 2011 – approx. **95%** annotated

Lucie Poláková (Mladová)

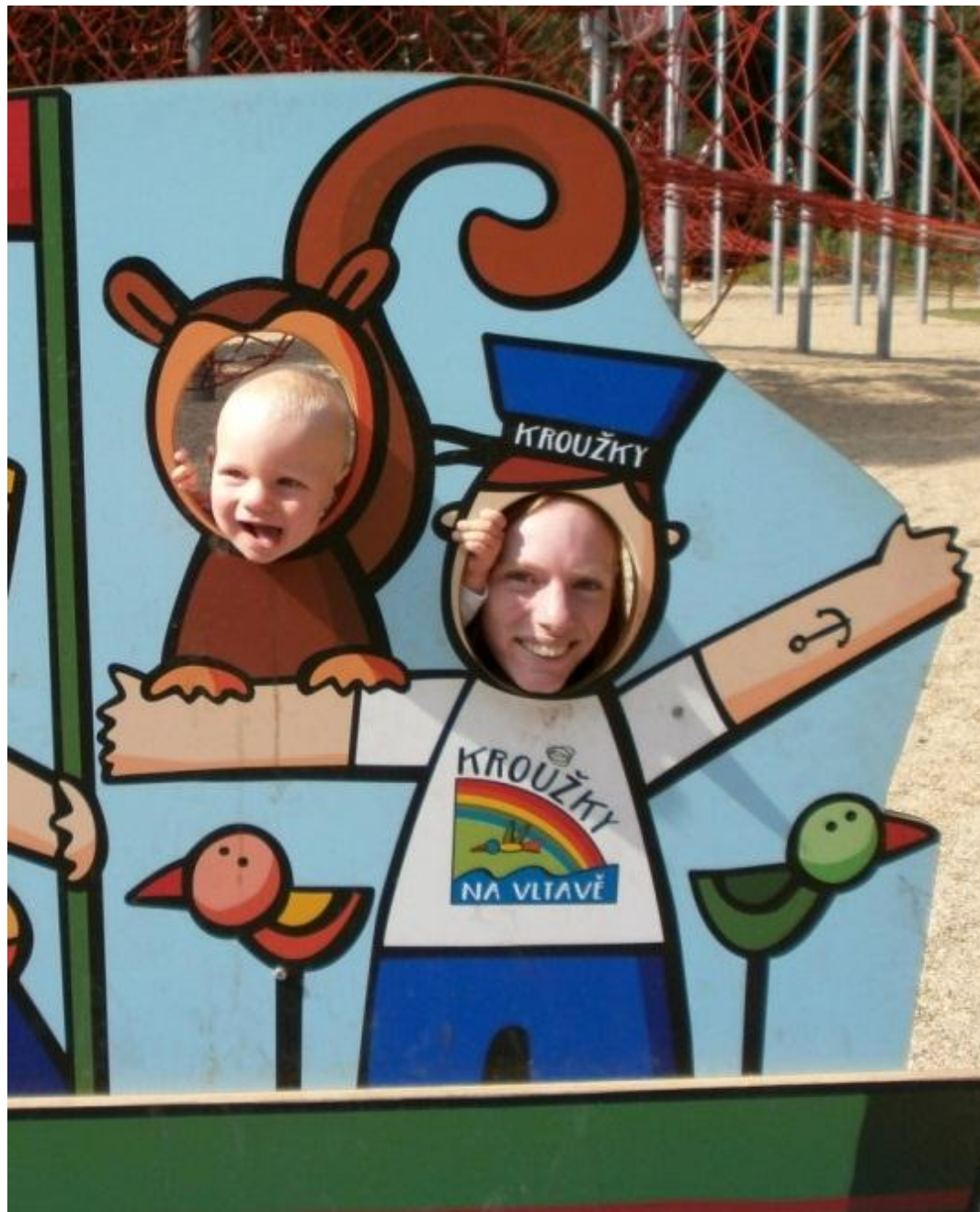
RECENTLY:

- May 2011: discourse workshop in Prague with Penn people
- July 2011: workshop *Annotation of Discourse in Large Corpora* in Birmingham (Bonnie Webber, Manfred Stede)
- December 12, 2011 Manfred Stede on Monday Seminar in Prague
- End of 2011 = publishing of sample data via UFAL (?)
 - = end of 3-year GAUK (Mladová)
 - = end of 3-year GAČR (Hajičová)

PhD: Dissertation topic:

The concept of discourse-level description for the PDT

Apart from that...



UFAL seminar
Michal Novák
15. 9. 2011, Malá Skála

School year 2010/11

- My 1st year of PhD studies
- Erasmus exchange programme
 - University of Primorska, Koper, Slovenia
 - 6 months
 - Interesting courses
 - Statistical Modelling
 - Algorithms on Graphs

School year 2010/11

- European Summer School of Logic, Language and Information 2011
 - Ljubljana, Slovenia
 - 2 weeks
 - Interesting courses
 - Roberto Navigli: Graphs in Natural Language Processing
 - Marjorie McShane and Sergei Nirenburg. Difficult Aspects of Reference Resolution: Description, Theory and Hands-On Practice

Research and Work

- Resolution of Czech NP coreference
 - 44% F-measure
 - DAARC 2011 conference, Faro, Portugal
- Utilization of anaphora in MT
 - Preliminary work
 - Error analysis
 - WDS 2011, Prague

Research and Work

- CzEng 1.0
 - Responsible for annotation of coreference
 - Porting coreference resolvers to TectoMT

Nino Peterek

- **Research**

- spoken data retrieval,
- automatic dialogue systems and prosody modelling,
- project Dialogy.Org, user interface to audio-visual corpora,
- development of corpus Dialog (in cooperation with AV ČR, ÚJČ),
- analysis of senior speech (in cooperation with AV ČR, PÚ).

- **Students**

- Automatic Topic Identification in Audio-visual Corpora, Spoken Data Retrieval (Mgr. P.Galuščáková),
- Community Speech Corpora Portals (Mgr. O.Krůza),
- Statistical Natural Language Processing Methods in Music Notation Analysis (Bc. J.Libovický).

- **Education**

- autumn course NPFL038 Fundamentals of speech recognition,
- spring course NPFL079 Algorithms for speech recognition.

On Restarting Automata, Functional Generative Description, and Mala Skala

Martin Plátek

KTIML, MFF UK, Prague

My father's family is from Mala Skala

Goals

- Our current effort is to study, and to present the linguistic know-how of the Prague linguistic group in the terms of analysis by reduction, restarting automata, and formal dependency structures. This know-how was developed since 60's under the name Functional Generative Description (FGD). I have started to co-operate in 70's.
- We work for/with two (essentially) different scientific aereas (groups of people):
 - 1) the area of formal languages and automata: M.Chytil, J.Vogel, F.Mráz, P. Jančar, R. Gramatovici, F. Otto, H.Messerschmidt, D.Pardubská, ..
 - 2) the aerea of computational and theoretical linguistics: J. Panevová, P.Sgall, K. Oliva, V. Kuboň, T. Holan, M. Lopatková, P. Pognan, ..

Formal frameworks for FGP

The original and the new version

- A (generative) serial composition of a CF-grammar, four push-down transducers and one finite transducer which gradually generates the individual levels of FGD

PS78 Plátek, M., and P. Sgall: A Scale of Context-Sensitive Languages: Applications to Natural Language. *Information and Control* Vol. **38**(1) (1978) 1–20

- A non-deterministic correctness preserving restarting automaton (analyzed are all four levels simultaneously)

LPS07 Lopatková, M., M. Plátek, and P. Sgall: Towards a formal model for functional generative description, analysis by reduction and restarting automata. *The Prague Bulletin of Mathematical Linguistics* Vol. **87** (2007) 7–26

Explicative power of FGD

Basic tasks

- L_w - the set of the correct (w -level) typographic sentences of the given language (**generative power**)
- L_t - the set of the meaning-representations of the sentences of the language (tectogrammatical representations)
- the **SH**-relation (**distinguishing relation**) between L_w a L_t describing the (gradual refinement of the) synonymy and ambiguity of the language;
- **a correctness preserving non-deterministic analysis by reduction** – uses a set of surface and tectogrammatical categories to ensure the correctness preservation of reductions

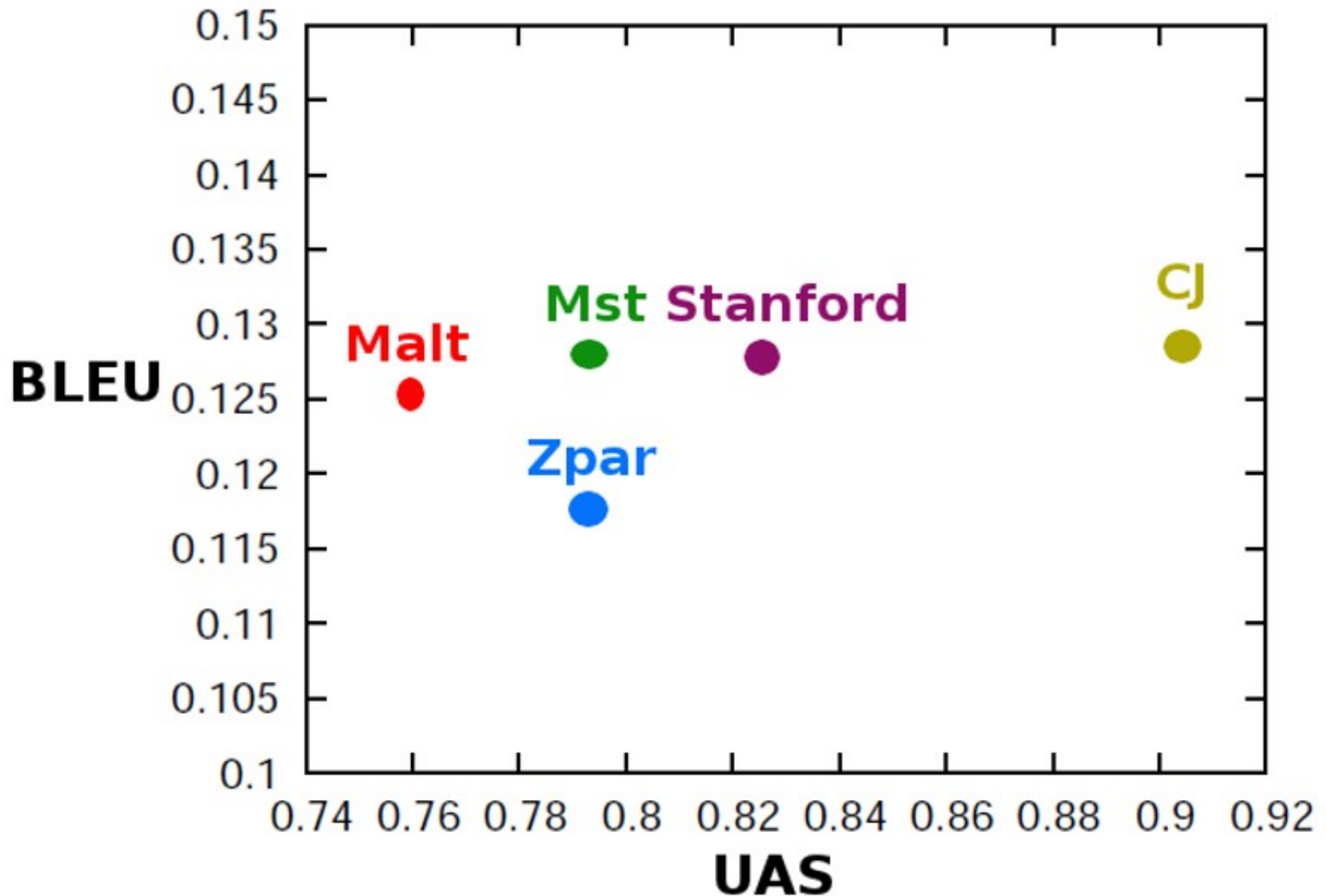
Summary and Future work

- The first types of restarting automata worked (with auxiliary symbols) in the sense of the standard formal language theory. Our effort to model the analysis by reduction in the FGD-sense lead us currently (roughly since 2005) to the notions of **characteristic and proper languages** in order to support more adequately the technique of lexicalization by restarting automata.
- **We plan to study the properties of SH-relations (formal translations) with the help of restarting automata, and their compositions.**
- To integrate free-order dependency grammars and restarting automata into a common paradigm. To use it for the study of the complexity and freeness of the word-order.
- **To learn properly to work with the positive and negative linguistic information simultaneously.**

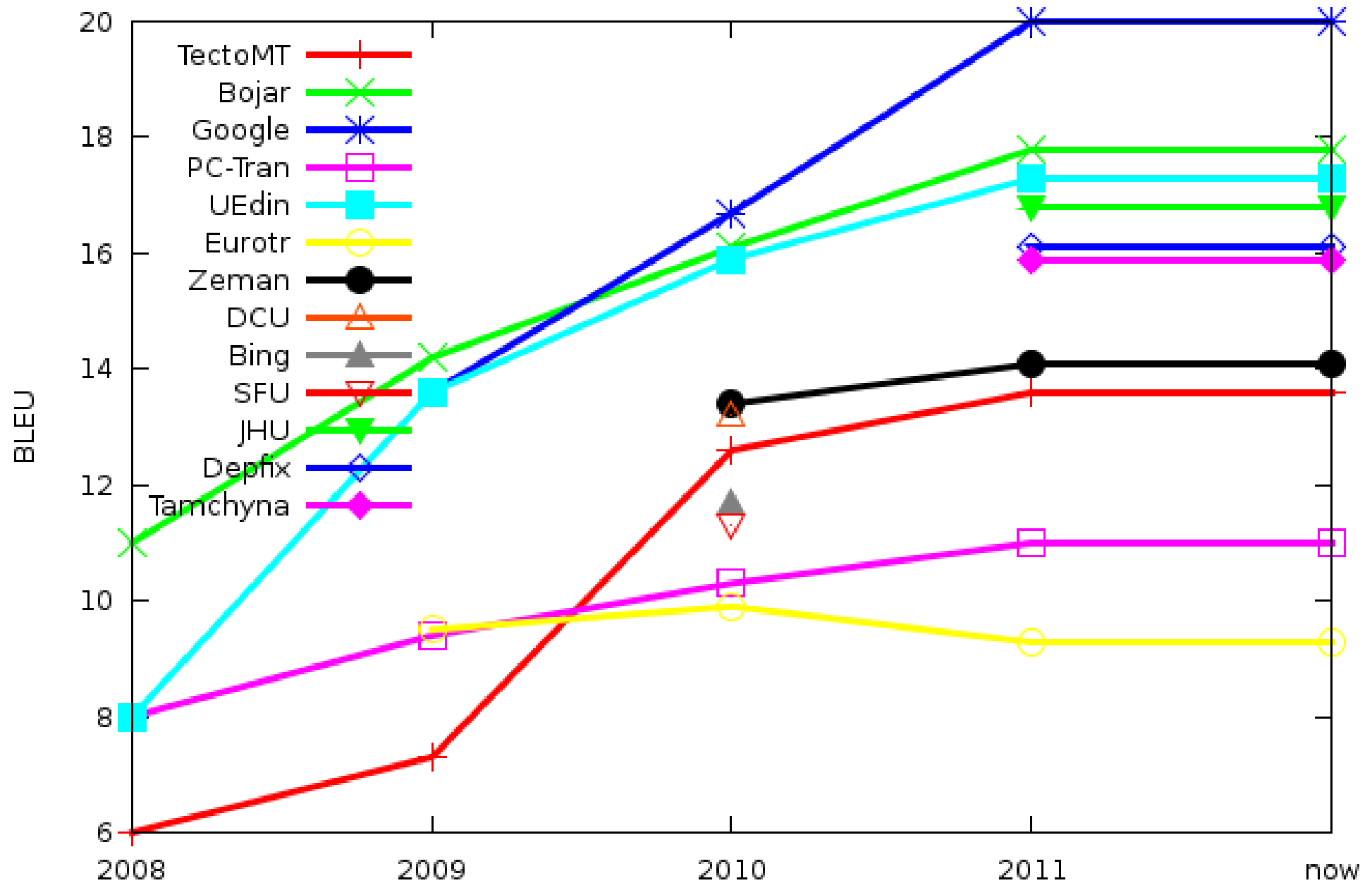
Martin Popel

- **Treex** framework
 - CPAN release, refactoring, testing etc (with Tomáš Kraut)
 - English analysis, integrating new tools (parsers)
- **TectoMT** machine translation (en-to-cs)
 - experiments with various parsers (with ZŽ, DM, NG)
 - post-processing – substitutions trained on ref-tst
- **PBML**
 - from tex to pdf, PBML style, mailing with authors
 - communication with printing house(s), Versita
 - for deadlines see <http://ufal.mff.cuni.cz/pbml> (Jan 8th)

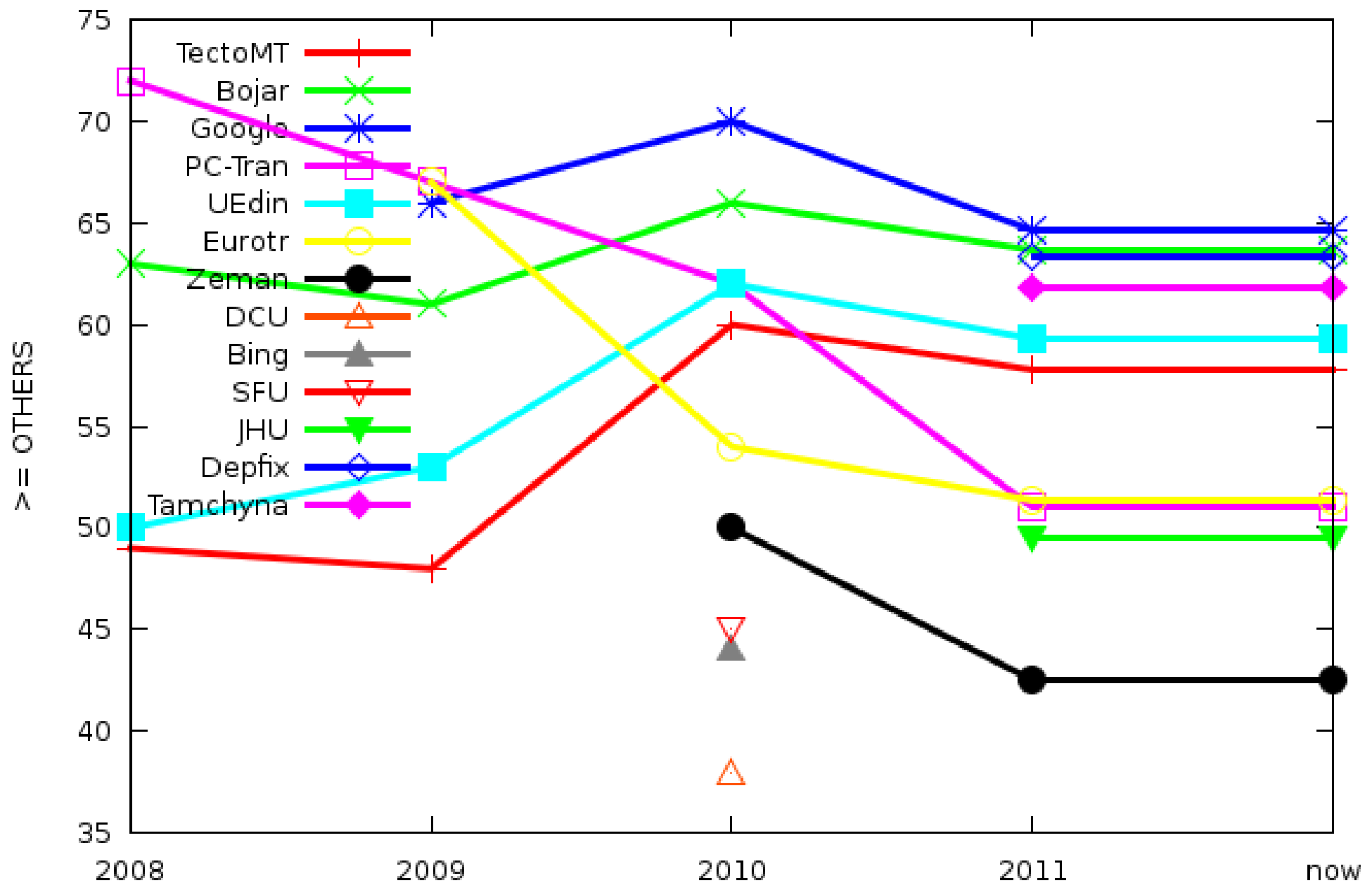
Extrinsic (BLEU) and intrinsic (UAS) evaluation of parsers



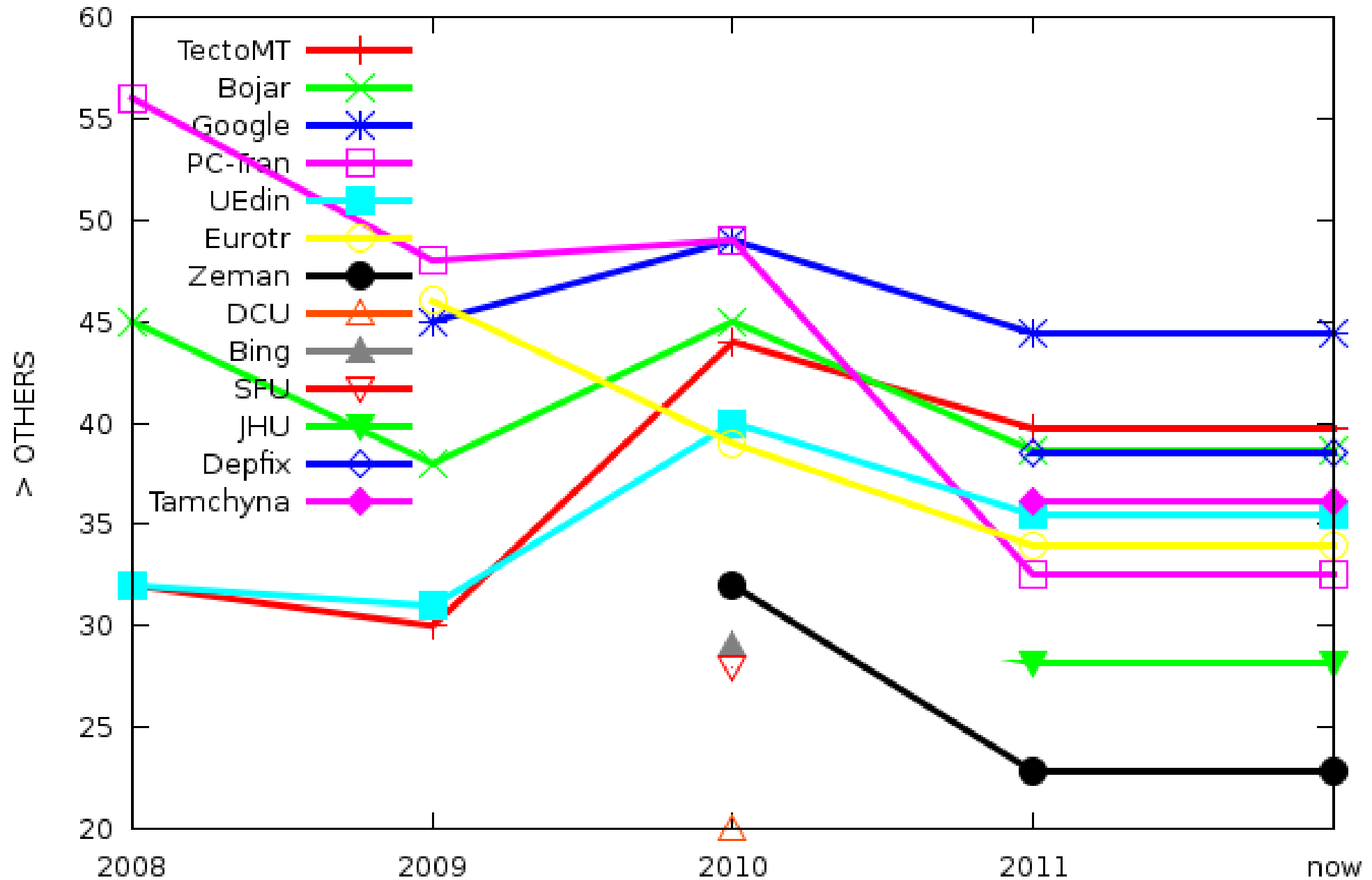
WMT: BLEU



WMT: manual ranking (\geq others)



WMT: manual ranking (> others)



Rudolf Rosa @ ÚFAL

- 23 years old
- lives in Kladno
- 1st year of Master studies
- `rosa@ufal.mff.cuni.cz`

First Steps...

- ◆ optimizing Treex code (former TectoMT) for speed and memory leaks
- ◆ doc. Ing. Zdeněk Žabokrtský, Ph.D.
- ◆ from February 2011 till April 2011

Currently...

- ◆ FAUST
- ◆ improving Statistical Machine Translation by grammatical post-processing using Treex
- ◆ Mgr. David Mareček
- ◆ since April 2011

Right Now

- ◆ MSTperl – reimplementing R. McDonald's MST parser in Perl as a part of Treex
- ◆ parallel parsing – parsing Czech sentence (SMT output) using features from its English counterpart (SMT input)

Future?

- ◆ FAUST (project till 2013)

Kateřina Rysov

Project

“From the structure of a sentence to textual relationships”

- GA 405/09/0729 (prof. Eva Hajičov et al.)
- linguistic research on information structure of a sentence
- the main topic: systemic ordering in Czech

Magdaléna Rysová

Project

“From the structure of a sentence to textual relationships“

- GA 405/09/0729 (prof. Eva Hajičová et al.)
- annotation of discourse relations
- topic: alternative lexicalizations of discourse connectives in Czech

Individual Report

Septina Dian Larasati

Mala Skala. September, 2011



Septina Dian Larasati

Indonesia

Undergraduate Degree : University of Indonesia, 2003-2007.

Master Degree : Free University of Bozen-Bolzano and
Charles University in Prague, 2008-2010.
Erasmus Mundus Language and Communication Technologies

PhD Study : Charles University in Prague, 2010 - Present.



Interest

Machine Translation, Language Resources, Morphology, Parsing...

septina@tilde.lv – larasati@ufal.mff.cuni.cz

CLARA

Host Partner



SIA Tilde
Vienības gatve 75a
LV-1004
Rīga, Latvia
tilde@tilde.lv

Work Package

Project 6C - Translation Tools and Resources for under-resource languages

Duration

August 1st, 2011 – April 30th, 2013

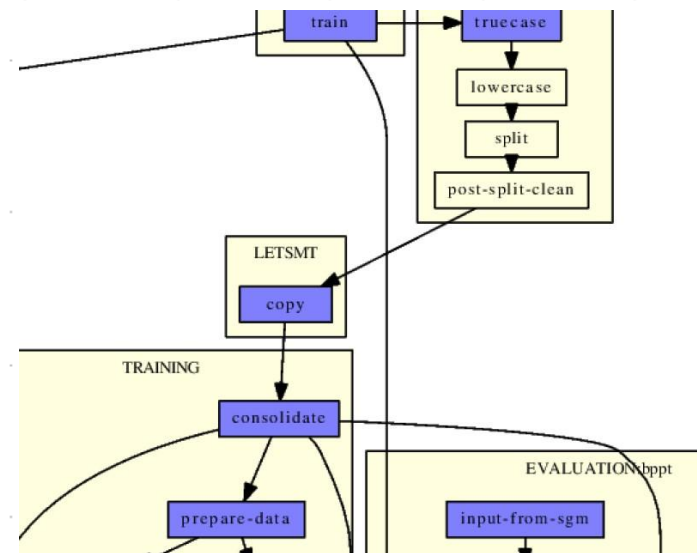
Current Project



LetsMT!
<https://www.letsmt.eu/>



Accurat
<http://www.accurat-project.eu/>



Moses EMS – Experiment Management System

Statistical Machine Translation

Indonesian – English

Language Tools:

- MorphInd: Morphological Tool for Indonesian
<http://ufal.ms.mff.cuni.cz/~larasati/MorphInd/MorphInd.html>

Language Resource:

- Indonesian – English Parallel Corpus

Ambitions:

Indonesian – English parallel Treebank

Indonesian Sentence Parser

New SMT reordering techniques

...

Thank you

Septina Dian Larasati

Mala Skala. September, 2011

CLARA



Magda Ševčíková

- involved in projects
 - GA ČR P406/2010/0875 – Prof. J. Panevová, “*Computational Linguistics: Explicit description of language and annotated data focused on Czech*”
 - MŠMT ČR LC536 – Prof. J. Hajič, “*Center of Computational Linguistics*”
 - post-doc project proposal on description of (selected) derivational relations in Czech
- teaching
 - course on academic writing, with Marie Mikulová
 - “*Professional Language and Style*”
 - 2004/2005–present
 - Faculty of Mathematics and Physics, Charles University in Prague, for master students
- involved in organization of Monday seminars / Fred Jelinek Seminar Series
 - **speakers and proposals welcome!!!**

- revision and refinement of tectogrammatical annotation – towards PDT 3.0 (2.5):

(i) pair/group meaning of Czech nouns

- plural forms of nouns *ruce* (hands/arms), *vlasy* (hair), *boty* (shoes), *sirky* (matches) etc. refer to a pair or a typical group of entities
 - cf. the plural form *ruce* denotes a pair or several pairs of hands/arms, not just many of them
- pair/group meaning considered as a grammaticalized meaning of Czech nouns
 - new grammateme *typgroup* (values *group*, *single*, *nr*)
- manual annotation of the pair/group meaning with selected nouns in the PDT 2.0 data, October 2010–January 2011, with Lenka Smejkalová and J. Panevová
 - by two annotators in parallel, inter-annotator agreement 75.1%, kappa 0.67
 - automatic assignment of the rest of the data
- publications
 - journal articles published in *Slovo a slovesnost* (72, 2011, pp. 163–176; with J. Panevová) and *PBML* (96, 2011, pp. 27–48; with J. Panevová and L. Smejkalová)
 - oral presentation at the ÚFAL *Seminar of Formal Linguistics*, March 7, 2011
 - conference paper (with J. Panevová and Z. Žabokrtský) published in the proceedings of the TLT 2010 workshop

(ii) verbal mood

- the grammateme *verbmod* used in PDT 2.0 data is to be substituted by the new grammateme *factmod*
- the *factmod* grammateme implemented by Jirka Mírovský

(iii) sentence modality

- in PDT 2.0 data, the sentence as a whole assigned a sentence modality (assertive, imperative, interrogative etc.)
- the revised proposal: sentence modality is to be specified for each clause in coordination structures, and for the root of title structures
- to be implemented by Jirka Mírovský

- further publications in 2010/2011

- conference paper on grammatical diatheses and other issues, published in the proceedings of Depling 2011 conference (with J. Panevová)
- conference paper on predicatives, to be published in the conference proceedings of the conference of the Institute for Czech Language
- oral presentations (with J. Panevová)
 - on corpora of Czech – for the Kruh přátel českého jazyka, March 9, 2011
 - on the revision of grammatememes for PDT 3.0 – at the *ÚFAL Seminar of Formal Linguistics*, November 1, 2010
 - on PDT annotation – for PhD students of linguistics at the Faculty of Philosophy and Arts, Charles University in Prague, October 21, 2010

Lenka Smejkalová

- Educational background
 - Student MFF UK 2004 – 2010
 - Bachelor's thesis: Document Similarity Visualization
 - supervisor Pavel Pecina
 - defended in September 2007
 - Master's thesis: Typical Usage Patterns of English Verbs
 - supervisor Martin Holub
 - defended in September 2010

Lenka Smejkalová

- Employee of ÚFAL since August 2010
- Research:
 - Lexical Semantics
 - with S. Cinková, M. Holub
 - Selectional Preferences
 - Process annotated data in PDEV, IAA
 - PDT
 - with J. Panevová, M. Ševčíková
 - prepare and handle the data for the annotation of number of nouns with pair/group meaning

Jan Štěpánek

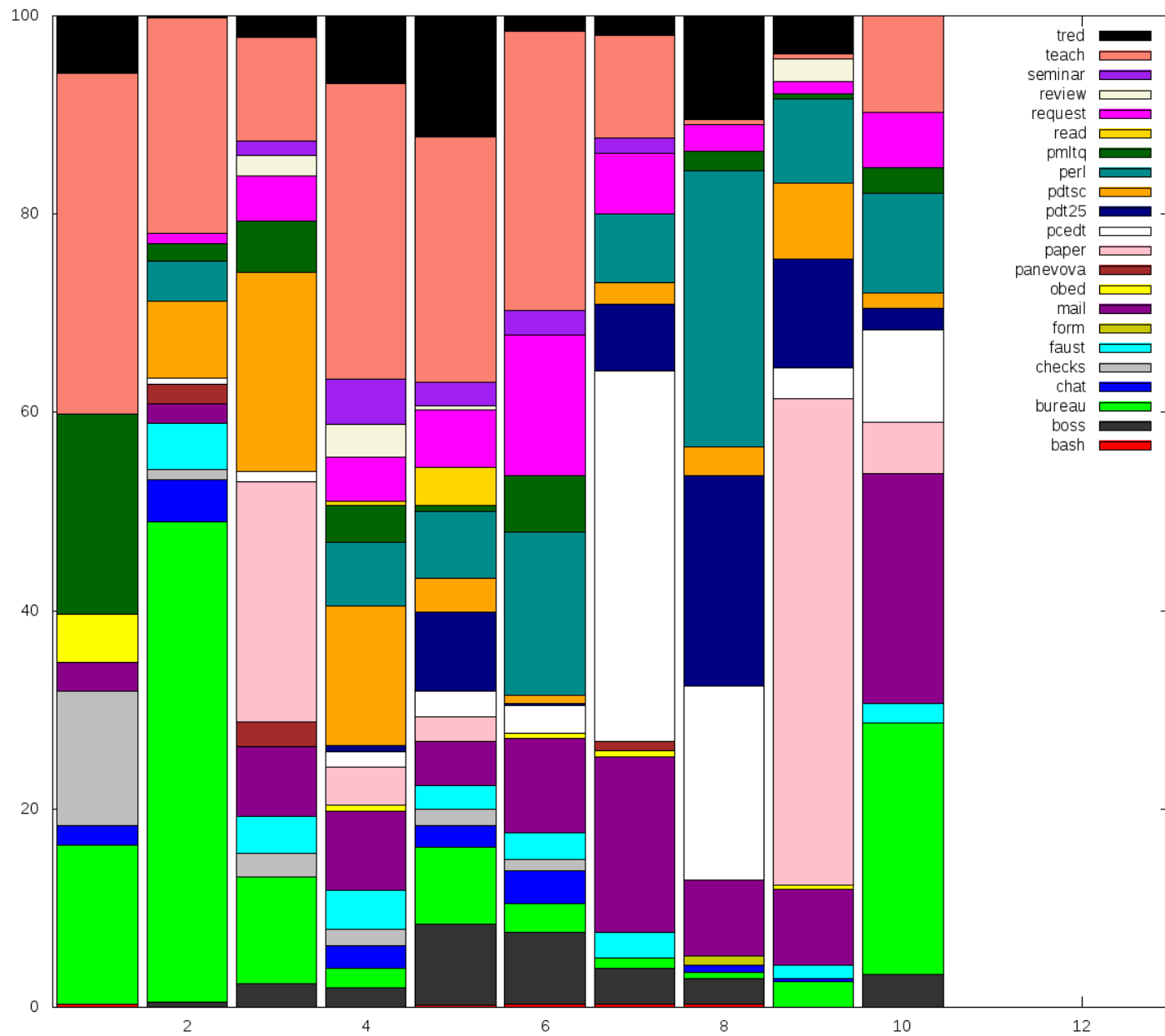
Paid from:

20% LINDAT

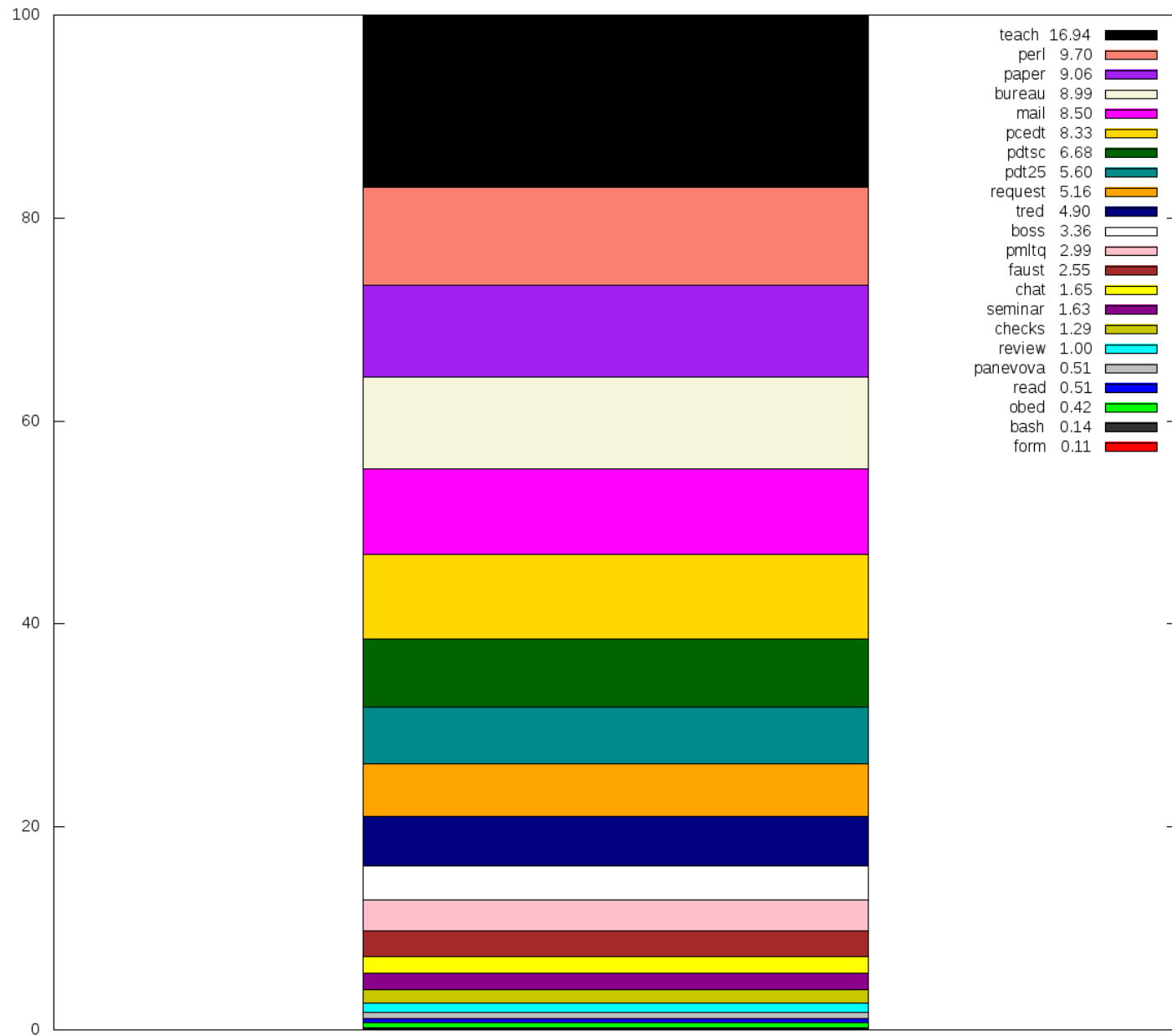
80% GAČR

- Tools for Revision and Tectogrammatical Annotation of a Czech Dependency Treebank

Real work



Real work



Pavel Straňák

- ✦ Multiword lexemes and NE for PDT 2.5 (with EB and JP)
- ✦ Scientific secretary of LINDAT
 - ✦ Clarin
 - ✦ a repository for data and tools (with KV and JM)
 - ✦ Michal Richter's Korektor
- ✦ EUDAT (with JH)
- ✦ Automatic processing of textual data (NPFL098)

Kateřina STUPARIČOVÁ

Education:

- Charles University in Prague, Faculty of Arts
Master's degrees: Information Science & New Media Studies
- University of Ljubljana, Slovenia, Faculty of Arts
Erasmus: Information Science and Librarianship

Main work experience:

NEWTON Media - media analyst (4 years)

Kateřina STUPARIČOVÁ

ÚFAL (since 08/2011):

European projects:

- CLARA
- T4ME (META-NET)
- FAUST
- Eurometrix Plus
- KHRESMOI
- EUDAT
- MosesCore

administrative, financial tasks

- r. 412
- tel. 4396
- email: stuparicova@ufal.mff.cuni.cz

Josef Toman

- 11/2006 – 06/2009: support for PEDT annotation
- 05/2009: Graduated at ÚFAL
- 07/2009 – 09/2010: somewhere else
- Since 10/2010: making ÚFAL's corpora ready to be published
 - PDTSE
 - PDTSC
 - PEDT 2.0
 - PCEDT 2.0
 - PDT 2.5

How to publish a corpus

- Get some data
- Write a web about them
- Create preview of the data for the web
- When needed, create TrEd visualisation
- Make sure it all fits to CD/DVD
- Write installers
- Write a script that does it all in one go
- Write a script for LDC so that they can check out that the previous script did a good job

What is it really about?

- Building a web of 100 000+ pages with Bash & Perl
- Transforming EngValLex into Javascript
- Storing 18 GB of SVG to DVD (no compression allowed)
- Getting very friendly with our cluster
- Writing scripts that write other scripts
- Exceeding the disk quota occasionally
- Knowing at least something about almost everything:
 - TrEd, (X)HTML, Javascript, MSI, CSS, XML, ICO, PML, XSLT, SVG, Treex, MEd, OGG, Perl, SVN, Flash, MP3, XSH, Gimp, Imagemagick, LRC, Bash, ...

Something to look at

- <http://ufal.ms.mff.cuni.cz/~toman/pdtse/>
- <http://ufal.ms.mff.cuni.cz/~toman/pdtsc/>
- <http://ufal.ms.mff.cuni.cz/~toman/pedt/>
- <http://ufal.ms.mff.cuni.cz/~toman/pcedt/>
- <http://ufal.ms.mff.cuni.cz/~toman/pdt/>

Zdeňka Urešová

- **GAUK:** Diathesis and transformations of surface expressions of valency arguments
(Topic for my PhD Thesis) – finished in July 2011 (E. Hajičová, J. Ptáček., J. Šindlerová)
 - Formal relation between verb with its arguments and their surface expression, especially in diatheses in PDT-Vallex
 - Czech verbs, from PDT 2.0 and extensions (WSJ translation)
 - New: transformation rules for creating form of diatheses from canonical (primary) valency frames
 - Purpose: natural language generation, consistency checking
- **PhD Thesis:** Valency of Verbs in the Prague Dependency Treebank
- Other projects:
 - **GAUK:** Multilingual source for verbal valency - finished in July 2011
(J. Šindlerová, M. Lopatková, O. Bojar, Z. Urešová)
 - **GAUK:** Enriching and Expansion of Lexicographic Resources- finished in July 2011
(E. Bejček, M. Lopatková, P. Straňák, K. Vandas, Z. Urešová)
- **Work on PDT-Vallex – WSJ translations**


Project Clarin (Common Language Resources and Technology Infrastructure)

- Eva Hajičová, Jan Hajič, Pavel Straňák, ...
- Webs: <http://www.clarin.eu>, <http://ufal-point.mff.cuni.cz>
- The aim of project is to unify the access to academic resources world-widely
- Used technologies: DSpace, Shibboleth, LDAP
- Position: Technical assistant (undefined :-))
- Responsibility:
 - Administration of server <http://ufal-point.mff.cuni.cz>
 - Cooperating with www.eduid.cz federation, creating user accounts
 - Administration of the test installation
 - Creating backups, setting used software, bug-fixing of used software
 - Administration and modifications of DSpace source code

Project Vallink

- Markéta Lopatková, Eduard Bejček, ...
- The aim of project is to unify VALLEX 2.5 and PDT-VALLEX vocabulary
- Used technologies: Perl, XML, XSLT
- Position: Technical assistant (undefined :-))
- Responsibility:
 - Transforming PDT-VALLEX to newly created SAMR data format
 - Tiny suggestions about „merging work-flow“
 - Developing of segments of source code
 - Developing of frames merging based on weighted regular expressions

Kateřina Veselovská

- ÚFAL since 2008
- GAUK 19008/2008 - A Multilingual Archive of Verbal Valency Characteristics (J. Šindlerová)
-  PCEDT 2.0
- PhD. Enriching the Treebank Annotation with Selected Phenomena from the Field of Pragmatics
 - in fact: sentiment analysis of PDT

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

*I **hate** grammatemes. x I **like** pancakes.*

- Current team:

Kateřina Veselovská, Jana Šindlerová, Jan Hajič jr., Mirek Týnovský
(special thanks to Ondřej Bojar & prof. Hajičová)

- Current state:

annotations running, rather satisfactory results

contributions mostly investigating the relationship between
polarity and negation

Other topics of interest

- Construction Grammar
- Lexical Semantics
- Text summarization

Zdeněk Žabokrtský (1/2)

- **research topics**

- **past**

- valency (VALLEX, Valency lexicon of Czech verbs)
 - treebanking (grammatemes and coreference in PDT)
 - anaphora resolution, parsing, named entities

- **current**

- using dependency syntax in machine translation
 - building and exploiting multilingual data (CzEng)
 - semi-supervised and unsupervised ML in NLP

Zdeněk Žabokrtský (2/2)

- **Courses taught in 2011/2012:**

- **Technology for NLP** (together with Jan Štěpánek)

- bash+perl+xml...

- **Language data resources**

- corpora, treebanks, lexical databases ...

- **Selected Problems in Machine Learning**



- one big adventure...

- **PhD students supervised in 2011/2012**

- Nguy Giang Linh, David Mareček, Martin Popel, Michal Novák, Nathan Green, Loganathan Ramasamy, Martin Majliš

Dan Zeman

- Hierarchical Machine Translation (Joshua)
 - preprocessing (word order transformation)
- Interset: Conversion of morphological tags from one tagset to another (both in Czech and cross-language)
 - Universal description of morphological tagsets
 - ISOCat, ISO TC37/SC3-4
- Multilingual Dependency Parsing
 - with Zdeněk, Martin, David, Hošť, Loganathan ...

Dan Zeman

- Parsing
 - Nivre's Malt Parser on Czech (UAS 86.08 %)
- Teaching
 - Disrupted: Computational NLP (but kept at ČVUT)
 - New course: Morphological and Syntactic Analysis
- “Dirty” (= non-scientific) work:
 - Bibliography maintenance (the “Biblio” database)
 - Address book maintenance (PBML, corpora registration...)

Vlád'a Kuboň

- Projects
 - Responsibility for the Res Informatica grant at UFAL,
 - Cooperation with companies (Hieronymus, Virtus)
- Research
 - Segmentation analysis
 - Segmentation of complex sentences into smaller linguistically motivated units and analyzing their mutual relationships
 - MT between related languages (Česílko)
 - Program committee of several workshops (EAMT, IIS, BSNLP etc.)
 - Project review for Islandic Research Agency
- Teaching
 - 2 lectures - „Úvod do poč. lingvistiky ...“ in Czech and in English, organization of the new NLP Applications lecture; 2 seminars for UFAL and 1 seminar of Automata and Grammars for CS students
 - Co-coordination of EM LCT Master program together with Markéta Lopatková,
 - UFAL secretary for teaching
 - Coordination of an Erasmus exchange with Saarbruecken, Koper and Tuebingen
 - Supervised 1 thesis (defended in January), 4 PhD. students (5 after October 1st), 1 BSc. thesis and 1 Term project

Ngũy Giang Linh

- **Coreference resolution in PDT**

- Doctoral study (Supervisor: assoc. prof. Ing. Zdeněk Žabokrtský, Ph.D.)
- PDT -> Treex:
 - Deletion detection (#PersPron, #Gen, #Unsp, #Cor, #QCor, #Rcp, #Benef)
 - Statistical grammatical and textual coreference resolution
- PCEDT:
 - Comparison of Czech and English coreference

- **Future Works**

- PDT -> Treex:
 - Completion of pronominal coreference resolution
 - Nominal and bridging anaphora resolution
- PCEDT:
 - Czech and English coreference resolution



Bushra Jawaid

Institute of Formal and Applied Linguistics
Charles University in Prague

Mala Skala, September, 2011

Past

- 2002-2006
 1. ***BS in Computer Science from University of Karachi:*** Final Project on Rule based English to Urdu Machine Translation.
- 2006-2008
 1. ***Worked as Software Engineer:*** Infinilogic pvt Ltd., Mixit Technologies pvt Ltd. Mostly worked using .net platform.
 2. ***Conference on Language and Technology CLT07:*** Compiled Bachelor's project and published paper.
- 2008-2010
 1. ***European Masters in Language and Communication Technologies (University of Malta and Charles University in Prague):*** Masters thesis on Statistical Machine Translation between Languages with Significant Word Order Differences.
 2. ***Conference on Language and Technology CLT09:*** Hindi to Urdu conversion beyond simple transliteration.

Present

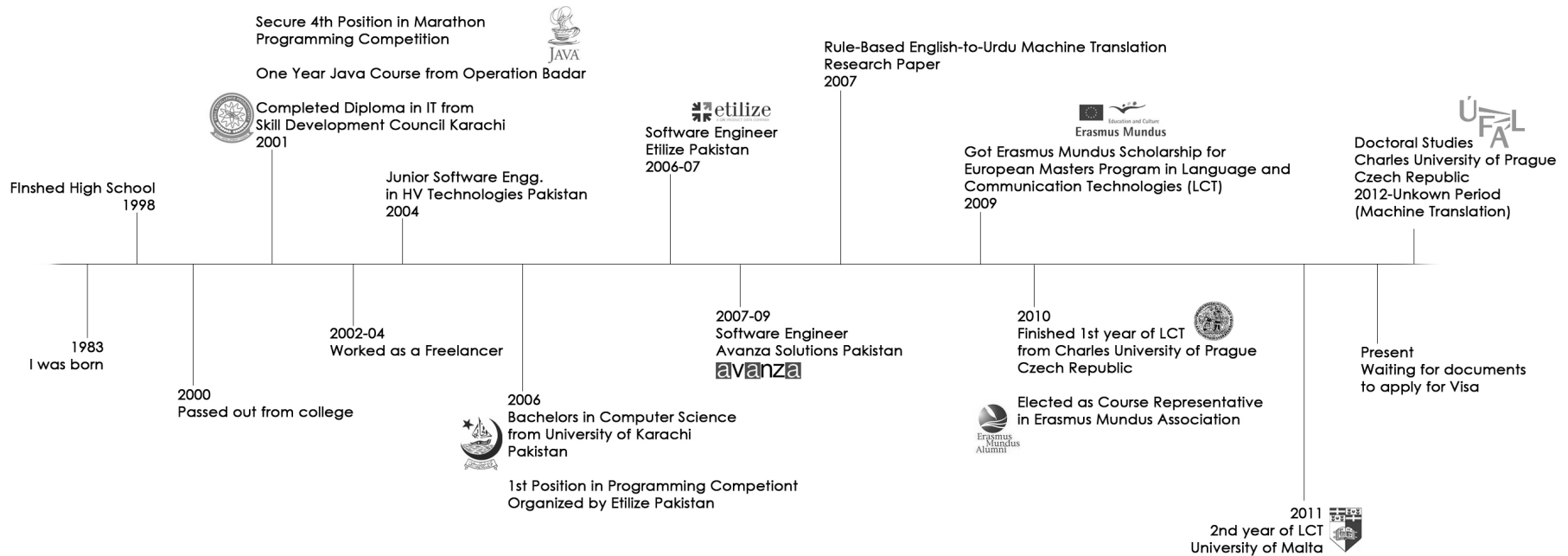
- 2010-Till Today
 1. PhD. On Statistical Machine Translation with Significant Word Reordering and Rich Target Side Morphology.

Amir Kamran

Institute of Formal and Applied Linguistics

Malá Skála, September 2011

Introduction



Future Plans and Interests

- Exploration of linguistic features in Machine Translation
- Applications of NLP (specially for mobile devices)
- Wants to do another masters in Photography
