Short presentations

ÚFAL seminar at Malá Skála
20. – 23. 9. 2012
Contents

Ondřej Bojar ................................................................. 3
Silvie Cinková .............................................................. 13
Ondřej Dusek ............................................................... 16
Nathan Green ............................................................. 18
Eva Hajicová ............................................................... 23
Barbora Vidová Hladká .................................................. 26
Pavlína Jínová .............................................................. 27
Filip Jurčíček ............................................................... 29
Filip Jurčíček: VYSTADIAL ............................................. 33
Václava Kettnerová ....................................................... 38
Veronika Kolárová ......................................................... 39
Veronika Kolárová: GACR .............................................. 40
Septina Dian Larasati .................................................. 44
Markéta Lopatková ....................................................... 47
Markéta Lopatková: Lexicographic Description of Syntactic and Semantic Properties of Czech Verbs .................. 50
David Mareček ............................................................ 54
David Mareček: FAUST .................................................. 55
Jirí Mírovský ............................................................... 59
Anja Nedoluzhko .......................................................... 64
Michal Novák .............................................................. 67
Jarmila Panevová .......................................................... 70
Pavel Pecina ............................................................... 76
Pavel Pecina: GACR ...................................................... 77
Pavel Pecina: Khresmoi .................................................. 82
Nino Peterek .............................................................. 88
Lucie Poláková (Mladová) ............................................. 89
Martin Popel .............................................................. 90
Loganathan Ramasamy ............................................... 91
Ondřej Bojar

- main interests:
  - MT and morphologically rich languages
  - manual and automatic MT evaluation
  - ML (for NLP)
  - dormant interests: parsing, valency, ..

- duties:
  - timesheets

- projects:
  - EM+ finished in May 2005, co-funded TectoMT and PCEDT 2.0
  - 4 running, 1 under consideration, 3-4 in planning
Running: MosesCore (Bojar)

- 2012-2014, EU FP7 CSA
- MT Marathons
  - funds for 5 students to MT Marathon yearly (suggest! only 1 went this year)
  - Sept 2014 in Prague; 2015 in Trento
- WMT shared task organization
  - My main interest: how to manually evaluate MT (maybe a discussion on pilot studies late tonight)
- Moses development
  - not our main duty but we (e.g. Aleš Tamchyna or Matouš Macháček) still participate
Running: GAČR Bojar

• 2010-2012, post-doctoral
• original goal: tight integration of phrase-based and deep-syntactic MT
• tangible outcomes in 2012:
  • experiments in targeting rich morphology
  • factored phrase-based models
  • towards automatic detailed error analysis
  • hopefully a chapter in a handbook (with Lori Levin)
Running: GAČR Zeman

• 2011-2013, „regular“ GAČR

• original goal: Czech in MT (with English, German, Spanish)

• my 2012 co-contribution:
  • hopefully experimenting with HyTERs for Czech
    – a technique to manually create thousands of correct reference translations
  • automatic detailed error analysis
  • hopefully some preliminary from supervising a Bc. thesis on German language modelling
Running: AMALACH (Hajič)

- 2012-2015, ministry of culture
- ÚFAL's task: MT of speech and short phrases for cross-lingual search in MALACH data
- my 2012 contribution:
  - nothing yet
  - I will be supervising Elena Manishina from Oct till Dec
Submitted: MT+CAT (Bojar)

- 2013-2015, TAČR Alfa
- With MemSource.cz
- Goal: tight integration of MT (Moses) into CAT
  - i.a. „MT auto-complete“
- Last time, the proposal was rejected for formal reasons.
Proposing: EMTEE (name will change)

- 2014-2016 EU FP7 STREP
- CUNI, UEDIN (P. Koehn), ISI (K. Knight), USTUTT (A. Fraser) + a company
- Foreseen goals: semantics in SMT, i.a.:
  - semantic language models (Kevin)
  - string-to-tree Moses with morph. modelling (Philipp)
    - Ondrej wants to apply Moses to t-layer
  - comparing t-layer with ISI's AMR (Ondrej)
  - verb frames and semantic roles in MT (Philipp)
Proposing: UTOPIA

- 2014-2016 EU FP7 STREP
- CUNI, SHEFF (L. Specia), ZURICH (M. Fishel), ? UPPSALA (S. Stymne) + a company
- Usable Translations Of Practically Anything
- Foreseen goals (negotiable):
  - automatic 'tiering' of output: ok/editable/rubbish
  - automatic post-editing like Depfix
  - error flagging to speed up manual post-editing
  - overall source and/or target coherence, e.g. terminology or anaphora
- Who is interested to join or learn more?
Proposing: TAMID

• 2014-2016 EU FP7 STREP
• CUNI, ?U. Valencia/Barcelona?, Pangeanic (M. Herranz), +some partners on semantics ???
• Truly Advanced Multilingual Information Discovery
• Foreseen goals (not quite clear to me yet):
  • some semantic parsing
  • search within semantic representation
  • MT at some point:
    – before parsing, MT of queries, pivoting or direct translations
  • my suggestion: search for keywords in other languages
    – What would be the French query for „Arab spring“?
• Who is interested to join or learn more?
Proposing: Post-Editing

- 2014-2016 EU FP7 STREP
- CUNI, Pangeanic (M. Herranz), ECI (European Captioning Institute), some Canadian film/live events/chat customers

Foreseen goals (not quite clear to me yet):
- starting point: translators hate to post-edit
- goal: make post-editing easier, more accepted, build community

Who is interested to join or learn more?
Silvie Cinková

RECENTLY FINISHED and to-be-done-soon WORK

• annotated corpora
  – PCEDT, PEDT
  – PDTSE

• minor revisions of the PEDT annotations
  – with Vojtěch Diatka and Eva Fučíková

• revision of the English TR documentation for a new release
CURRENT WORK

• lexical description of English verbs
  (with Martin Holub, Lenka Smejkalová and Vincent Kríž)
  – multiple annotation of collocation preferences of verbs, based on Corpus Pattern Analysis (P. Hanks): lexical resource „VPS-30-En“

• automatic collocation extraction (with Lenka Smejkalová)
  – PMLTQ queries that identify relevant collocate positions of verbs in analytical trees
  – combination of dependency and word-order rules
Silvie Cinková

WORK IN PREPARATION

(Large-scale analysis of multimodal data, Czech Science Foundation Centre of Excellence with the Technical University, Pilsen and MU)

• with Pavel Pecina and Martin Holub
• text-in-the-wild annotation
• similarity search in the ProfiMedia image database, based on visual similarity vs. similarity of surrounding text (captions etc.)
FAUST – MT output improvement (till 1/2013)

- Fixing grammar in Czech MT outputs by analyzing and generating back
- My focus is Czech deep NLG, but I work also on:
  - Formemes (morpho-syntactic function annotation)
  - Automatic functor assignment
  - English-Czech translation in TectoMT
  - Treex NLP framework

VYSTADIAL – spoken dialog systems (from 10/2012)

- Natural language generation within a spoken dialog system
- Stress on naturalness, re-usability for other languages/systems
Ondřej Dušek

Study

- Master’s in math. linguistics at ÚFAL from 2010
- Starting Ph.D. study at ÚFAL now
  - Supervisor: Filip Jurčíček
  - Topic: NLG in spoken dialog systems
- Finishing master’s in German philology at the Faculty of Arts

Other Interests

- Machine learning
- Valency
- Languages in general, particularly English, German, Dutch
# Background

<table>
<thead>
<tr>
<th>Activity</th>
<th>Organization</th>
<th>Dates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Undergraduate</td>
<td>North Carolina State University</td>
<td>1999 - 2003</td>
</tr>
<tr>
<td>Internship</td>
<td>IBM</td>
<td>2000</td>
</tr>
<tr>
<td>Masters</td>
<td>George Washington University</td>
<td>2003 - 2006</td>
</tr>
<tr>
<td>Fulbright Fellowship</td>
<td>University of Iceland</td>
<td>2006 - 2007</td>
</tr>
<tr>
<td>Graduate Research</td>
<td>North Carolina State University</td>
<td>2008 - 2010</td>
</tr>
<tr>
<td>PhD</td>
<td>Charles University/ CLARA</td>
<td>2010 - ...</td>
</tr>
</tbody>
</table>
Dependency Parsing

- Nathan Green and Zdeněk Žabokrtský, Hybrid Combination of Constituency and Dependency Trees into an Ensemble Dependency Parser. Innovative hybrid approaches to the processing of textual data, Workshop EACL, Avignon, France, pp. 19-26, 2012

- Nathan Green, Loganathan Ramasamy and Zdeněk Žabokrtský, Using an SVM Ensemble System for Improved Tamil Dependency Parsing. ACL Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages, Jeju, Republic of Korea, 2012

Collaborative Environments


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
Training

Clara Training Activities

- Prague - New developments in CL

Conferences

- ACL - Jeju
- EACL- Avignon
- LREC – Turkey
- Paclic - Bali
Future Research

- Continue looking at dependency parsing and its influence on statistical machine translation
- Application of semi-supervised techniques to ensemble/hybrid dependency parsing
Eva Hajičová

- Participant of the LINDAT (PI: Jan Hajič) and GAČR (PI: Š. Zikánová) grants
- Other responsibilities: The Prague Bulletin of Mathematical Linguistics (helped by Martin Popel and others) and „Monday“ seminar (helped by Aleš Tamchyná)
• Organizational responsibilities:

• **Workshop on Discourse:** COLING 2012, Mumbai – **Deadline Sept. 30, 2012**

• **Dependency Linguistics** (DepLing) – Prague, August 27 – 30, 2013 (related workshop: Meaning and Text Theory, August 30-31, 2013), **DEADLINE April 15, 2013**
• Main research interests:
• **Discourse** relations (incl. topic-focus articulation – PhD: Magdalena Rysová), coreference relations, annotation of discourse relations – PhD: Lucie Poláková-Mladová, sentiment analysis – PhD: Kateřina Veselovská)
• (Linguistic aspects of) **Dependency** syntax
Barbora Vidová Hladká

- **Research**
  - INTLIB
    - A joint TAČR project of Sysnet Ltd and MFF UK (KSI, ÚFAL)
    - Document processing (mainly searching) – NLP+LinkedData
    - Legislative documents
  - GAČR – Šárka Zikánová
    - „Koreference, diskurs a aktuální členění v kontrastivním pohledu“
    - Salience assignment
  - An exercise book of Czech morphology and syntax STYX
    - Čapek: a new component
  - Lgame: games with a purpose
    - iOS implementation

- **Teaching**
  - Introduction to Machine Learning (in Computational Linguistics) with Martin Holub (in other words, Machine learning #1, for #2 and #3 see Zdeněk Žabokrtský)

- **Student supervision**
  - Vincent Kríž: a PhD student from Fall 2012
Pavlína Jínová

- Ph.D. student of Czech language at Faculty of Arts
- teaching at Faculty of Arts
- since 2009 discourse annotation project at UFAL (lead by prof. Hajičová, Lucie Poláková and Šárka Zikánová)

> discourse annotation = marking discourse relations (e.g. conjunction, reason-result) expressed between text spans containing finite verbs and signaled by surface-present connective (e.g. and, because, therefore)
Two parts of discourse annotation:

- manual part: all inter-sentential relations and those discourse intra-sentential relations whose tectogrammatical representation do not convey discourse semantics accurately (e.g. functor ADVS divided into more specific types, clausal APPS semantically classified)

- automatic part: all intra-sentential relations which correspond exactly to their tectogrammatical counterparts

- manual annotation finished in January 2012, checking procedures finished in June (7705 relations)

- automatic annotation run in July, checking procedures finished this week (10 482 relations)

- data will be published (hopefully) in November 2012
Filip Jurčíček

• Main activities in the last year
  • Teaching:
    – STATISTICAL DIALOGUE SYSTEMS
    – Master theses
  • Proposal: 2012 FRVS teaching grant
    – Introduction into Bayesian techniques in machine learning
  • Project: VYSTADIAL
Teaching – a new course

- STATISTICAL DIALOGUE SYSTEMS
  - Summer semester 2012
  - Lectures - 2 hours per week
  - 6 students
  - 4 students finished the course

- The course will be updated in 2013
  - Lectures and practicals - 2 + 1 hours per week
  - Using internally developed SDS
Master theses

- David Marecek
  - Bayesian inference for belief tracking is SDS

- Jan Hajic Jr.
  - Discriminative models for belief tracking in SDS
2012 FRVS teaching grant

- Invited two post-doc from Machine Learning Group at Cambridge University
- To give 8 lectures on
  - Approximate Inference: sampling methods, variational Bayes and expectation propagation.
  - Non-parametric Bayesian Methods: Gaussian processes and Dirichlet processes.
  - Bayesian Sparsity: spike and slab priors, dependency in sparsity enforcing priors, group sparsity.
  - Bayesian Latent Variable Methods: probabilistic matrix factorizations,
  - Bayesian mixtures of Gaussians.
- + 6 practicals
Filip Jurčíček

• New project: VYSTADIAL
  • Development of statistical methods for spoken dialogue systems

• Funding for 4 PhD students
  – 3 students are already hired (as of 17.9.2012)
  – the last student will be (hopefully) hired the next year
Students

• New students:
  – Ondrej Dusek – an UFAL student
    • you probably know him
    • will work on natural language generation
  – Matej Korvas – a formal Bc. UFAL student
    • Erasmus Mundus student – Saarbrucken and Melbourne
    • will work on spoken language understanding
  – Lukas Zilka – a FIT Brno student
    • spent 6 months at Open University at Milton Keynes, UK
    • will work on dialogue state representation for complex dialogue domains
Building UFAL dialogue system #1

- **ASR**
  - HTK + OpenJulius ASR decoder
  - training scripts finished and the decoder is already integrated
  - collecting ENG and CZ telephone data
  - English:
    - still using UCAM data - 40 hours of audio
    - getting approximately the same WER as in CAM
    - collecting my own EN data using MTURK
  - Czech:
    - so far collected 8 hours of transcribed speech
    - available 20 hours of un-transcribed speech (CZSwitchboard, Repeat after me)
    - WER is promising ;-) 
- **VAD**
  - implemented using GMM model trained on force-aligned speech
Building UFAL dialogue system #2

- **SLU**
  - implemented using dialogue act classifiers
  - now only working on 1-best hypotheses
  - awaiting modification for decoding from N-best lists and confusion networks

- **DM** – a baseline DM is not implemented yet

- **NLG**
  - Ondrej Dusek started on Czech NLG

- **TTS**
  - English - using free Festival/Flite
  - Czech – using ZCU/Speechtech – voice Iva
Test domains

- **TownInfo**
  - giving tourist information about a specific city, e.g. bars, hotels, restaurants

- **Prague Airport**
  - giving information about flight arrivals and departures

- **Planned participation in**
  - Dialog State Tracking Challenge
  - Jan – March 2013
Kettnerová Václava

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

Research interest
• valency of verbs, diatheses, alternations (thesis on lexical-semantic conversions in Czech – the similar relations btw. syntactic structures *The farmers loaded hay onto the truck* – *The farmers loaded the truck with hay* – and their representation in the valency lexicon VALLEX) defended in June 2012), lexical-semantic representation, semantic classification of verbs
• Now: **light verbs**: btw. nouns and verbs; their lexicographic representation and detection
• In future: derivational relations of verbs with respect to changes in valency
Veronika Kolářová

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

Participation in two GAČR projects:

- **Systematic, economical and corpus-based description of valency properties of Czech deverbal nouns (theory and practice)**
  - P406/12/P190
  - post-doctoral project, 2012-2014
  - principal investigator

- **Computational Linguistics: Explicit description of language and annotated data focused on Czech**
  - P406/10/0875
  - standard project, headed by prof. Panevová; 2010-2013
  - team member

Main topics of interest:

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

GA ČR P406/12/P190

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

Theory:

• To complete the description of adnominal counterparts of adverbal objects expressed by prepositionless cases
  – In past (dissertation): Dative, Accusative
    • dodávky zákazníkům ‘delivery-NOM.PL customer-DAT.PL’
    • varování řidičům / řidičů ‘warning-NOM.SG driver-DAT.PL / driver-GEN.PL’
  – Now: Genitive, Instrumental
    • jeho dotek puku ‘he-POSS.SG touch-NOM.SG puck-GEN.SG’
    • nákaza chřipkou ‘infection-NOM.SG flu-INS.SG’

• To specify unique valency properties of some semantically compact groups of nouns
  – nouns of saying, nouns of giving, nouns denoting mental state or dispositions

Practice:

• To incorporate knowledge about nominal valency into
  – the treatment of dictionary entries (PDT-VALLEX)
  – the guidelines for annotation of PDT (not annotation as such)
Current work I.

• Specific shifts in surface forms of participants and their consequences for the meaning of the noun and for the treatment in the lexicon
  – Does the specific shift imply the change of the meaning of the noun?
    • e.g. typical shift ACC → GEN vs. specific shift ACC → PS
    • lov velryb.PAT rybáři.ACT
      ‘hunt-NOM.SG whale-GEN.PL fisherman-INS.PL’
    • lov na velryby.PAT *rybáři.ACT
      ‘hunt-NOM.SG on whale-ACC.PL fisherman-INS.PL’
    • občasný lov na velryby.PAT
      ‘occasional-NOM.SG hunt-NOM.SG on whale-ACC.PL’
  – Treatment in the lexicon
    • lov₁ ‘hunt’: ACT(2,poss,7) PAT(2,poss)
    • lov₂ ‘hunt’: ACT(2,poss) PAT(‘on’+4)
  – SLE conference (Stockholm, August 2012)
Current work II.

- Issues of valency of nouns derived from verbs with a participant expressed by prepositionless Genitive
  - Adnominal counterparts of adverbal prepositionless Genitive
    - Also some nouns derived from verbs by non-productive means keep $\text{Gen}_{\text{Adnom}} \leftrightarrow \text{Gen}_{\text{Adverb}}$
    - *odvaha spolupráce* ‘courage-NOM.SG cooperation-GEN.SG’
    - *dotyk míče / s míčem* ‘touch-NOM.SG ball-GEN.SG / with ball-INS.SG’
    - Conference „Čeština v pohledu synchronním a diachronním“ (Prague, June 2011)
  - Nominal constructions with two participants expressed by prepositionless Genitive
    - *domáhání se Ireny Riškové ACT svých peněž. PAT*
    - ‘claiming-NOM.SG REFL Irena-GEN.SG Rišková-GEN.SG she-REFL.POSS.GEN.PL money-GEN.PL_TANT’
    - Conference “Grammar and Corpora” (Prague, November 2012)
  - Possessive adjective or pronoun as the correlate of Gen$_{\text{Adnom}}$ corresponding to Gen$_{\text{Adverb}}$
    - *dotknutí se nervu. PAT / jeho. PAT dotknutí se*
    - ‘touching-NOM.SG REFL nerve-GEN.SG / it-POSS.SG touching-NOM.SG REFL’
    - Conference “Word and form in the structure and communication” (Bratislava, December 2012)
Septina Dian Larasati
Indonesian

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. (Statistical Machine Translation id-en)

Currently under the CLARA grant (August 2011-April 2013)
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

September 2012. Mala Skala, Czech Republic
Google Summer of Code (GSoC) 2012 Mentor: Apertium id-ms
Raymond Hendy Susanto
National University of Singapore

Linguistic Resources:
- MorphInd : Indonesian Morphological Analyzer
- IDENTIC : Indonesian-English Parallel Corpus

Publications:
2011

2012
Thank you

Find more about me @ http://ufal.mff.cuni.cz/~larasati/
Markéta Lopatková

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)
"Teaching projects":

- EM LCT (Language and Communication Technologies) together with Vladislav Kuboň
  3 students for 2012-13
  - new phase: 2013-2018?
    selected for funding

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

- involved in a preparation of BSc. "General Computer Science" in English (from 2013/14)
Markéta Lopatková (3)

**Teaching:**

– Mathematical analysis  
  winter + summer term, a practical course, BSc.  
– Prague Dependency Treebank  
  with Honza Štěpánek  
– Mathematical Methods in Linguistics  
  ???

**Supervising:**

– 4 PhD students, 3 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
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)
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)
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
     • model of a lexicon
  4. Branching Programs
Central Funding

- **PROVOZ** (teaching money)
  - ca 1.7 mil. CZK salaries (3.7 full contracts)
  - 540 th. CZK other costs
- **PRVOUK** (research money)
  - ca 3.1 mil. salaries (5.8 full contracts)
  - 500 th. CZK other costs
- **Specific Research (?)**
  - 130 th. CZK other costs
David Mareček

- Grants:
  - FAUST - Feedback Analysis for User adaptive Statistical Translation
- My research topic:
  - Unsupervised Dependency Parsing
- Other:
  - Treex, word-alignment, tecto-alignment, parsing, …
- Teaching:
  - Selected problems in Machine Learning (with ZŽ)
- Supervision:
  - Master student Rudolf Rosa
FAUST
Feedback Analysis for User adaptive Statistical Translation

- FP7 European Union project
- February 2010 -- January 2013
- Project partners:
  - University of Cambridge
  - Universitat Politecnica de Catalunya
  - Charles University in Prague
  - Language Weaver
  - Softissimo
- Our team:
  - Jan Hajič, David Mareček, Ondřej Dušek, Rudolf Rosa
FAUST
Feedback Analysis for User adaptive Statistical Translation

3 Goals:

- To provide our experimental MT systems online and let other users to use it and to provide feedback (labs.reverso.net)
- To develop mechanisms for incorporating user feedback into the commercial MT engines and provide real-time adaptation to the feedback (reverso.net)
- To integrate natural language generation into MT to improve translation fluency and reduce negative feedback from users
Two our systems are running on labs.reverso.net:
- TectoMT (Popel et al.) EN->CZ
- Moses (Bojar et al.) EN->CZ, CZ->EN

Problem:
- nobody uses it => we have almost no feedback from users

Plan:
- to ask at least UFAL people to use it and to provide some feedback
  - The annotations are planned on October-November

DEMO
FAUST
Feedback Analysis for User adaptive Statistical Translation

- DEPFIX system (developed by Rudolf Rosa)
  - automatic correction of phrase-based MT Czech outputs
  - correcting agreement (Subj-Pred, Adj-Noun, Prep-Noun)
  - WMT 2012 winner (Google + Depfix > Google)

EN source sentence ➔
CS translation dependency tree ➔
Improved translated CS sentence

CS translated sentence
tagging & parsing using also word-alignment and features from EN
apply morphological agreement rules and regenerate changed words
Annotation of Discourse

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

- annotation tool (an extension of TrEd), data management, measurement of the inter-annotator agreement
- data mining, data corrections
- (semi-)automatic annotation of the intra-sentential relations
- publication of the data
Prof. Panevová, Magda Ševčíková, and others

• sentence modality (titles, coordinated clauses instead of coordinations)
• grammatememes (factmod, tense, diatgram)
• in PDT 2.x svn
Annotation of Anaphora

Anja Nedoluzhko, and others

- annotation tool (an extension of TrEd), data management, measurement of the inter-annotator agreement
- published in December last year (over PDT 2.0)
- to be published along with discourse (over PDT 2.5)
- annotation of 1st and 2nd person
Public Service

- purchase and management of software (but not SW from Microsoft)
- purchase and management of data published by LDC
- assistant management of Amoeba (database of employees at ÚFAL)
Anja Nedoluzhko

coreference annotation at ÚFAL

- PDT 2.5 = ready data (web, CD 2011, to be published together with discourse in 2012)
- PEDT – annotating nominal coreference (without bridging) – 5 annotators
- PDT 2.5 for 1. and 2. person
- PCEDT for 1. and 2. person
• Collaboration with the discourse team
• comparing coreference annotation results in different languages
• anaphora resolution competition in Russian
• With Jarka – regular word-formative patterns in Czech and Russian, planned to be detailed from the linguistic and technical point of view
Anja Nedoluzhko

Variability of languages in time and space / Variabilita jazyků v čase a prostoru
Magda Ševčíková, Šárka Zikánová, Anja Nedoluzhko – starting in October 2012

- Languages of the world, possible classification (e.g. language families, number of speakers)
- Overview of phonemic inventories and writing systems, problems of transliteration and transcription
- Linguistic typology: using phonological, morphological and syntactical aspects and their combination (Skalička’s typology)
- Language universals: absolute, statistical and implicational universals
- Influence of diachronic language processes on the language variability
- The relation between spoken and written forms of a language
- Speakers’ attitudes to language variants
Research and work:

- CzEng 1.0
  - Automatic coreference annotation
  - Segmentation of the CzEng data into shorter chunks of texts to retain as much coherence as possible (eventually rejected for release) - with Ondra Bojar and Zdeněk Žabokrtský

- TectoMT
  - WMT12 competition
  - Improved translation models trained on the CzEng 1.0 data
  - with Zdeněk Žabokrtský, Martin Popel, Martin Majliš, Ondra Dušek and David Mareček
- Research and work:
  - Identification of the type of “it”
    - Anaphoric / non-anaphoric / pleonastic
    - Combination of statistical system NADA (Bergsma and Yarowsky, 2011) with manually designed rules
    - Accuracy improvement 3.5%
    - with Nguy Giang Linh and Katka Veselovská
  - Context-based lexical translation models
    - Improve the lexical choice (e.g. “bank”) using the information from the context – wider than just surrounding words (surface, in the syntactic tree)
    - For TectoMT
    - Work in progress
Research and work:

- Cached models in MT
  - Use the cache of previously translated words (phrases) to suggest the lexical choice of the word being translated
  - For Moses
  - With Liane Guillou, Nicola Bertoldi, Robert Grabowski, Sorin Slavescu and Jose de Souza
  - Work in pause

- Treex development
  - Framework for anaphora resolution (or other binary relations)
  - Compact maxent translation models
  - Step by step building a framework for ML

Researchers: Panevová 0,4
Bejček 0,5
Cinková 0,3
Hlaváčová 0,3
Kettnerová 0,5
Klujeva 0,2
Kolářová 0,3
Mikulová 0,6
Mírovský 0,7
Nedolužko 0,2
Rysová M. 0,4
Smejkalová 0,5
Ševčíková 0,1
2012:
Annotation of newly introduced grammatemes:
- pair/group meaning with selected nouns – introduced in PDT 2.5 (Ševčíková, Smejkalová)
- verbal grammatemes (rezultative1 vs. pasivum (Panevová, Mírovský), rezultative21, rezultative22 – also in PDTSC (Hlaváčová, Panevová)
  factmod, sentmod – Ševčíková, Mírovský
- Continuation of tectogrammatical annotating of PDTSL (Mikulová)
- PCEDT – published (Cinková)
Workshop on the grant project – April 12th, 2012
with a programme:

- Eduard Bejček, A play with phrasemesis - a pleasure for the mental capacity. Multiword expressions in PDT.
- Jan Štěpánek, On further improvement of PDT 2.0 a 2.5 data.
- Jiří Mírovský, On annotation of intersentential relations in PDT.
- Jarmila Panevová, A new schema of verbal grammatemes and its reflection in the resultative meaning.
- Magda Ševčíková, Annotation of pair/group meaning in the PDT data. A revision of modal meanings: the grammateme
- of sentence modality.
Publications


Publications - continuation


Pavel Pecina

- assistant professor since Jan 2012
- previously: post-doc researcher at Dublin City University, Ireland
- projects:
  - **Khresmoi** – Medical information analysis and retrieval. EU FP7-ICT Collaborative Project, 2010-2014
  - **CEMDI** – Center for Large-Scale Multi-modal Data Interpretation, GACR Excellence in Basic Research, 2012-2018
- research interests
  - statistical machine translation, domain adaptation
  - cross-language information retrieval, image retrieval, speech retrieval
  - other NLP tasks (spelling correction, multiword expressions, etc.)
- teaching:
  - New (Master/PhD) course on **Information Retrieval** starting in October
CEMDI: Center for Large Scale Multimodal Data Interpretation

**Funding:** Czech Science Foundation (GAČR), Excellence in Basic Research

**Code:** P103/12/G084

**Budget:** 14 Million CZK

**Duration:** 7 Years, Jan 2012 - Dec 2019

**Consortium:** 4 Institutions

1. ČVUT: Czech Technical University, Prague, **Prof. Jiří Matas**
2. MU: Masaryk University, Brno, Prof. Pavel Zezula
3. ZČU: University of West Bohemia, Pilsen, Prof. Josef Psutka
4. UK: Charles University in Prague, Pavel Pecina

**People:**

- Silvie Cinková, Martin Holub, Pavel Pecina
- Lenka Smejkalová, Anna Vernerová, Ema Krejčová
Projects goals

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

Experiteese

- ČVUT: image (video, pictures) processing (text in the wild)
- ZČU: speech processing (spoken archives, subtitling)
- MU: information retrieval, similarity search (images, text)
- UK: natural language processing, machine translation
Tasks: Reading text in the wild

Given an (arbitrary) picture with some text, what is the correct word order/text reading?
Tasks: Recommendation of illustrative images

Given a large collection of (annotated) photographs and a text of an article, what are the best illustrative pictures?

Domy budou muset mít energetické štítky, i přes Klausovo veto

19. září 2012 11:38

Majitele domů čeká nová povinnost. Budou muset zajistit, aby se na domech objevily štítky informující o jejich energetické náročnosti. Zákon prošel přes veto prezidenta Václava Klause, který ho oznácel z nejškodlivějších a nejzbytečnějších, protože podle něj bude znamenat další náklady pro majitele domů.

Žena v noci vstávala do práce, v pracovně našla zloděje

18. září 2012 14:17

Zloděje v pracovně nachytala po svém nočním probuzení osmapadesátiletá žena ve Zlině. Lupič do bytu vnikl oknem otevřeným na ventilaci, z místa stihl uprchnout. Policie po neznámém mladém muži párá.
Advertisement

- We are looking for prospective students interested in these and other related topics
- Possible colaboration with ČVUT, ZČU, MUNI
- Contact: pecina@ufal.mff.cuni.cz
Khresmoi: Multilingual and multimodal access to biomedical information (Jan Hajič)

**Funding:** EU FP7, Collaborative project, 14 Million EUR, 4 Years

**Consortium:** 12 Institutions

- University of Applied Sciences of Western Switzerland; Atos Spain, Madrid, Spain; Charles University, Prague, Czech Republic; Dublin City University, Ireland; University of Duisburg-Essen, Germany; Evaluations and Language resources Distribution Agency, Paris, France; Health on the Net, Geneva, Switzerland; Vienna University of Technology, Austria; Medical University of Vienna, Austria; Ontotext, Sofia, Bulgaria; The University of Sheffield GATE team, UK; Society of Physicians in Vienna, Austria.

**CUNI people:**

- Jan Hajič, Jarka Hlaváčová, Zdeňka Urešová, Jakub Bystroň, Pavel Pecina
Projects goals

- Information extraction from unstructured or semi-structured biomedical texts and images
- Linking the extracted information to information in knowledge bases
- Automated estimation of the level of trust
- Automated analysis and indexing for medical images in 2D (X-Rays), 3D (MRI, CT), and 4D (MRI with a time component)
- Adaptive user interfaces to assist in formulating queries and display search results via ergonomic and interactive visualizations
- Support of cross-language search, including multi-lingual queries, and returning machine-translated summaries
UFAL tasks

- Support of cross-language search, including multi-lingual queries, and returning machine-translated summaries
- Statistical Machine Translation adapted to the medical domain
- Languages: English, Czech, French, German
- Translation of non-English search queries to English
- Translation of English summaries to other languages
- Translation realized as a (distributed) webservice running on a computer cluster
Welcome to Khresmoi

Khresmoi makes searching for medical related topics easier, faster and more successful.

Some of the great features of this release include:

- Search the MIMIR index
- Get suggestions to expand your query
- Get spelling corrections
- Personal library for registered users
- Many other small improvements
- For developers: improved modularization

Stay tuned for further amazing features to come!
Description

Recurrent corneal dystrophy. Reticular opacity in the superficial cornea.
Translation Webservice Architecture

HTTP Interface

Application Server
HTTP Web Service

Port TCP/5001
XMLRPC Interface

Worker 1

Worker 2

Worker 3

Worker 4

Worker 5

Worker 6

Sentence splitter
Moses EN-DE
Tokenizer
Recaser

Shuffle

HTTP Interface

http://quest.ms.mff.cuni.cz:8888/khresmoi
Projects

• Tools for on-line speech corpora creation and exploration with Petra Galuščáksová and Oldřich Krůza (web services development for LINDAT).
• Speech corpora collection (ROMI, AMALACH).

Teaching

• Fundamentals of speech recognition and generation NPFL038 (2/1).
• Natural computing for learning and optimisation NPFL107 (2/1).
Lucie Poláková (Mladová)

**Current project:**
Annotation of discourse structure on PDT (Czech texts, based on tectogrammatics)

- Together with: Eva Hajičová, Šárka Zikánová, Jirka Mírovský, Pavlína Jínová
- Cooperation with UPenn and prof. Aravind Joshi’s team
- Grant support: till 2011 3-year GAUK (Mladová), 3-year GAČR (Hajičová), recently: GAČR (Zikánová, “Interplays” till 2015), KONTAKT (Hajičová, till 2012)

**Recent about the project:**
PENN: May 2012, Discourse Workshop in Penn (reply to our Prague workshop a year ago)
POTSDAM: November 2012, Invitation to Jena MULDICO workshop on Multilingual databases and corpora of connectives (reply to invitation of prof. Stede to Monday seminar in December 2011)
COST: New huge project on multilingual corpora with linked connectives

**PhD: Dissertation topic:**
The concept of discourse-level description for the PDT
Martin Popel

- **Treex** framework (tutorial for CLARA 2012)
- **TectoMT** machine translation (CzEng 1.0 for WMT12)
- Improving **HamleDT**, transforming coordinations
- **PBML** (next deadline January 8th 2013)
- Technical reports (2012 deadline December 1st)
- Teaching
  - Modern Methods in CL "**Reading group**"
  - (with ZŽ in summer: Language Data Resources)
- Supervising
  - 3 bachelor students
    (Michal Koutný: Word prediction using language models)
Translation Tools and Resources for Under-Resourced Languages

Loganathan Ramasamy

CUNI

Sep 20, 2012
Outline

1. Work so far
   - Treebanking
   - Parallel Corpora
   - Morphological Segmentation

2. Publications
Treebanking - TamilTB

- Tamil Dependency Treebank - TamilTB
  - Developed a small PDT style treebank for Tamil.
  - Annotation layers: (i) m-layer and (ii) a-layer.
  - POS tagset size: 234 unique tags
  - Number of dependency relations: 21
  - Data size: 600 sentences from news corpus.
TamilTB - Availability

- Data and documentation available at: http://ufal.mff.cuni.cz/~ramasamy/tamiltb/0.1/
- Also available for search in INESS system http://iness.uib.no/iness/treebanks
Parallel Corpora: English-Tamil

- Freely available corpora:
  - 38K (Matt Post et. al., WMT 2012)
- Corpus collection through web sources

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Size (sentences)</th>
<th>Words/sen (en)</th>
<th>Words/sen (ta)</th>
</tr>
</thead>
<tbody>
<tr>
<td>News</td>
<td>121369</td>
<td>26.5</td>
<td>19.6</td>
</tr>
<tr>
<td>Cinema</td>
<td>39544</td>
<td>15.3</td>
<td>10.2</td>
</tr>
<tr>
<td>Bible</td>
<td>30871</td>
<td>25.4</td>
<td>13.4</td>
</tr>
<tr>
<td>Total</td>
<td><strong>191784</strong></td>
<td><strong>24</strong></td>
<td><strong>16.7</strong></td>
</tr>
</tbody>
</table>

- Challenges:
  - Parallel corpora (news and cinema) can be noisy due to automatic alignment.
  - Transliteration errors.
- The data will be released soon.
Unsupervised Morphological Segmentation

- Uses Bayesian approach (Gibbs sampling).
- Focuses on agglutinative languages (Tamil and Telugu).
- Datasets available: 1500 words (Tamil) and 998 words (Telugu).
2012: Publications & Talks


Jana Straková
graduate student („Natural language and the human brain“)

March – August 2012: maternity leave

August 2012:

… coming back from September (25% from LINDAT-CLARIN)
Pavel Straňák

一. Lexical semantics: multiword expressions, named entities

一. MWE in PDT 2.5

二. Teaching "intro to NLP and data" for humanities' students

三. Weakness for slightly exotic languages

四. LINDAT Centre

五. European Projects Clarin and EUDAT
Magda Ševčíková

• April to October 2012 – on maternity leave
• involved in projects
  – GA ČR P406/12/P175 “Selected derivational relations for automatic processing of Czech”
    • post-doc project, 2012–2014
    • principal investigator
    • project led by Jarmila Panevová, 2010–2013
    • team member
• teaching
  – course on academic writing, with Marie Mikulová / Veronika Kolářová
    • “Professional language and style”, for master students, Faculty of Mathematics and Physics
    • 2004/2005–present
  – new courses, Autumn 2012
    • 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

sevcikova@ufal.mff.cuni.cz

Malá Skála, September 2012
Current work:

- on the topics of the post-doc project:
  - deadjectival derivates in Czech, currently focusing on so-called predicatives
    - e.g. ráno bylo deštivo ‘in the morning it was rainy’
    - existence of counterparts with the suffix -e (ráno vypadalo deštivě ‘the morning looked rainy’),
      problem of the part-of-speech categorization of the predicatives, syntactic functions of the
      predicatives and their counterparts etc.
  - conference paper on predicatives (just being) published in the proceedings of the conference
    Čeština v pohledu synchronním a diachronním
  - two papers accepted for linguistic conferences in Prague and Bratislava (November and
    December 2012)

- on the revision and refinement of the tectogrammatical annotation:
  (i) pair/group meaning of Czech nouns (with Jarmila Panevová and Lenka Smejkalová)
    - assigned to the PDT data in 2010/2011, published as a part of PDT 2.5
    - manual annotation of the pair/group meaning within the PDTSC data in 2011/2012
  (ii) verbal mood (with Jiří Mírovský)
    - the grammateme verbmod (used in the PDT 2.0 data) was replaced by the new grammateme
      factmod, the grammateme tense had to be changed in many cases as well
    - to be published as a part of PDT 3.0
(iii) sentence modality (with Jiří Mírovský)

- in the PDT 2.0 data: sentence modality (assertive, imperative, interrogative etc.) specified
  - for the root node of each sentence (of each tree), i.e. a sentence consisting of several coordinated clauses was assigned a single sentence modality,
  - for the root node of a direct speech subtree,
  - for the root node of a parenthesis subtree
- revised proposal: sentence modality specified for each clause in coordination structures, and (in addition to direct speech and parentheses) for the root node of title structures
- to be published as a part of PDT 3.0
- conference paper with Jiří Mírovský (“Sentence Modality Assignment in the Prague Dependency Treebank”) at the 15th International Conference on Text, Speech and Dialogue (TSD 2012) in Brno, September 2012

„Nevšímejte si jich, jsou to blbci,” řekl mi plynulou angličtinou jeden z přihlížejících. (“Do not take notice of them, they are idiots,” told me one of the onlookers in fluent English.)
GA ČR P406/12/P175

Selected derivational relations for automatic processing of Czech

- principal investigator: Magda Ševčíková
- post-doc project
- 2012–2014
- budget: app. CZK 500k for each year

- description of selected word-formation relations in Czech, namely of the relations between adjectives and their derivates
  - main goal: (i) to contribute to the theoretical research on word-formation, using corpus data, and (ii) to revise and broaden the lexical-semantic annotation of PDT
  - some partial tasks of the project:
    - to classify deadjectival derivates as syntactic derivates or lexical derivates (according to Kuryłowicz and Dokulil)
      - syntactic derivates have the same meaning as their base adjectives but differ in syntactic functions
      - lexical derivates differ from the base adjectives in meaning
    - to reflect this classification within the tectogrammatical annotation of Czech data
      - syntactic derivates can be represented by the t-lemma of the base adjective
    - to build a database of deadjectival derivates
Jana Šindlerová

- CZENGVALLEX
  - Currently with Eva Fučíková and Zdeňka Urešová
  - Interlinking of PDT-Vallex and Engvallex on the basis of parallel tecto-trees annotation
- New version of TrEd extension ready, new course of data annotation just beginning
- Expected full data release: 12/2013
- Funded by LINDAT-CLARIN
Jana Šindlerová
Jana Šindlerová

- SENTIMENT ANALYSIS of Czech data
  - Currently with Katka Veselovská, Jan Hajič, jr. and Mirek Týnovský
  - Funded by GAUK 3537/2011
Jan Štěpánek

Paid by:

GAČR (75%)
Tools for Revision and Tectogrammatical Annotation of a Czech Dependency Treebank

PRVOUK (25%)
Teaching

- Winter: NPFL092
  - NLP Technology (with Zdeněk Žabokrtský)

- Summer: NPFL075
  - Prague Dependency Treebank (with Markéta Lopatková)
### Responsible for

<table>
<thead>
<tr>
<th>TrEd</th>
<th>PML-TQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tree Editor</td>
<td>Query Language and Engine</td>
</tr>
<tr>
<td>OS independent</td>
<td>Several Servers and Clients</td>
</tr>
</tbody>
</table>
Ask me if you want to know something about

- Perl
- bash
- XML
- PML (Prague Markup Language)
Aleš Tamchyna

- Ph.D. student, 1st year.
- Advisor: RNDr. Ondřej Bojar, Ph.D.

Research:
- Statistical (phrase-based) MT.
- Participated in 2012 CLSP JHU Workshop
  - Domain Adaptation in Machine Translation.
  - My part: mainly Moses hacking.
- Just defended my Master thesis:
  Feature Selection for Factored Phrase-Based MT.

Other:
- Helping with ÚFAL Monday seminar organization.
Zdeňka Urešová

• **Kreshmoi Project** with Pavel Pecina, Jakub Bystroň, Jarka Hlaváčová
  – Topic: Medical Information on the Internet for general public and professionals (PI: JH)
  – User Evaluation preparation for the user test cases
  – Query translation for testing (mostly medical terms)
  – Preparation of general MT test data

• **PDTSC** with Jan Hajič, Jan Štěpánek, Marie Mikulová + annotators
  – Project topic: Part of the PDTSL (spoken language annotation) project (Č is Czech)
  – Extension of the valency lexicon PDT-Vallex of Czech verbs

• **Postdoc Proposal (GAČR)**
  – A comparison of Czech and English verbal valency based on corpus material (theory and practice)
  – Description of verbal valency in Czech and English
  – Description of interlinking of translational verbal equivalents
  – Data preparation together with Jana Vejvodová, Eva Fučíková

• **Intelligent library (TAČR Project)** Barbora Hladká & …
  – Analytical annotation (in preparation)
Zdeňka Urešová

GAČR POSTDOC PROJECT

(PROPOSAL)

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

A comparison of Czech and English verbal valency based on corpus material (theory and practice)
1st February 2013 + 3 years

• **Goals:**

  – To describe the relation between Czech and English valency frames

  – To build a Czech-English Valency Lexicon with explicitly linked verbal senses and their arguments/adjuncts

  – A comparative description of the argument structure of translation equivalents
A Cross-linguistic Comparison of Valency Behavior of Czech and English Verbs

- **Theoretical comparative studies** focused on differences in Czech and English verbal valency structure
  - a description of verbal valency in both languages
  - a description of interlinking of translational verbal equivalents with drawing a follow-up comparison between the achieved results
  - based on the valency theory of the FGD and on its application to the PDT
  - a specification of relations of verbal valency frames in both languages, relating to PDT’s semantic and morphosyntactic levels

- **Plus hands-on experience of work with corpus data**
  - The Czech-English valency lexicon (PDT-Vallex and EngVallex) will be interlinked at the level of verb arguments, as well as linked to the data (Prague Czech-English Dependency Treebank).
Kateřina Veselovská

- ÚFAL since 2008, currently at the conference:-/
- Ph.D. student
  "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*

• Current team:
  Kateřina Veselovská, Jana Šindlerová, Jan Hajič jr., Mirek Týnovský (supervisors prof. Hajičová & Ondřej Bojar)

• Current state:
  - sentiment-annotated corpus SubLex1.0 (5000 lemmas)
  - several polarity classifiers with rather satisfactory results
  - implementation of SubLex to TrEd (in progress)
  - annotation guidelines (in progress)
Other topics of interest

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

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

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

• Parsing
  – Nivre’s Malt Parser on Czech (UAS 86.08 %)

• Teaching
  – Morphological and Syntactic Analysis
  – Disrupted: Computational NLP (but kept at ČVUT)
  – New course next summer semester: “New Language”

• “Dirty” (non-scientific) work
  – Bibliography maintenance (the “Biblio” database)
  – Address book maintenance (PBML, corpora registration…)
GAČR P406/11/1499

• Titled *Czech in the Machine Translation Era (CZECHMATE)*
  – Dan Zeman
  – Ondřej Bojar
• Non-English translation (e.g. Czech-German)
• Phrase-based translation
• Named entities
• Pivot languages
Paraphrases from Moses Phrase Table
(MTM project)

Dan Zeman
Eva Hasler
Christian Buck
Why and How

• Multiple reference translations of one text
  – Useful bat scarce
  – Could be simulated by paraphrases?

• Bilingual phrase table
  – \( \text{English}_1 \) corresponds to \( \text{German}_i \)
  – \( \text{English}_2 \) corresponds to \( \text{German}_i \)
  – \( \text{English}_2 \) paraphrases \( \text{English}_1 \)
Corpora Used

● News Commentary + Europarl
● Language combinations: en/cs – cs/en/de/fr/es
● Bitext sizes:
  – from 650K sentence pairs (cs-*)
  – to over 2M sentence pairs (fr-en)
● Phrase table sizes:
  – from ca 50M lines
  – to ca 120M lines
What's Done

• Have not got beyond very simple heuristics

• Simple matching:
  – en1 → fr
  – en2 → fr
  – => paraphrase(en1, en2)

• Plus filtering

• No computation of probabilities yet
Filtering Uninteresting Stuff

● We may want to keep some of that in real application
● But it is not interesting to look at now
● Thus discard paraphrase if…
Filtering Uninteresting Stuff

- No letters (just digits and/or punctuation)
- Sole difference is
  - In upper/lowercase: European Parliament ↔ European parliament
  - In non-letters: Croatia and Turkey ↔ Croatia , and Turkey
- Substring (probably both extracted from one sentence): the Czech ↔ of the Czech
Substrings

• Should only consider whole words
  – Currently also discards *Czech ↔ Czechoslovak*
    • This one is not a good paraphrase but it should probably be discarded by a different technique. What about *Srinivas B ↔ Srinivas Bangalore*?
  – Even with whole words, this will disappear :-(
    • *Chavez ↔ Mr Chavez*

• Extend to non-continuous substrings?
  – This could also come from one sentence:
    • *Churchill said ↔ Churchill once said*
Stopwords

• We currently require more difference than just punctuation
• Same with stopwords? (Not implemented yet)
• Consider this example:
  – are 40 ↔ the 40 ↔ from 40 ↔ just 40 ↔ about 40
    ↔ of EUR ↔ to EUR ↔ up to 40 ↔ and for 40 ↔
    to 40 will ↔ at least 40 ↔ more than 40
Large Para Lists Are Untrustworthy

Mr ↔ Sino  Asia ↔ India drawn a China ↔ the PRC ↔ s China ↔ t China ↔ China by ↔ China up ↔ a matter ↔ be China ↔ by China ↔ China and ↔ China had ↔ China may ↔ China not ↔ China now ↔ China off ↔ China the ↔ China was ↔ China who ↔ countries ↔ for China ↔ not China ↔ win China ↔ , China for ↔ China , and ↔ China . FDI ↔ China does ↔ China ends ↔ China ever ↔ China from ↔ China held ↔ China must … and many more!
Results?

- Almost 6M equivalence classes (English via Spanish)
- Then intersect various lang pairs => better precision?
- E.g. intersect English paraphrases from
  - en – cs – en
  - en – de – en
  - en – es – en
  - en – fr – en
Good Output Examples

- I applaud this initiative ↔ I support this initiative ↔ I welcome this initiative
- Chen Guancheng ↔ Chen Guangcheng
- governments are trying to ↔ governments are attempting to
Less Good…

- Mr Babitsky ↔ he
- the EU wants ↔ the EU does not want to

- (One might be tempted to add manually the EU does not know what they want)
Šárka Zikánová

• Discourse
  – Organizational and financial questions of projects concerning discourse and coreference (GAČR, Kontakt, new Kontakt application)

• Lectures
  – Variability of languages in time and space (together with Magda Ševčíková and Anja Nedoluzhko)
  – Information structure and discourse analysis
  – (together with prof. Hajičová)
Coreference, discourse relations and information structure in a contrastive perspective

• 2012-2015
• Šárka Zikánová, prof. Hajičová, Anja Nedoluzhko, Barbora Vidová Hladká, Kateřina Rysová, Lucie Poláková
• Cooperation with Magdaléna Rysová, Pavlína Jínová, Jiří Mírovský (prof. Panevová´s GAČR, Lindat)
• 6.9 mil. CZK
GAČR P406/12/0658
Coreference, discourse relations and information structure in a contrastive perspective

• Interplay of coreference, discourse and information structure (Czech, English, German)
  – Contextual boundness and coreferential chains
  – Textual conditions of surface ellipsis (e.g. Information structure)
  – Means indicating a break in the discourse
  – Alternative lexicalizations of discourse connectives (*the reason is...*)
Kontakt ME10018
Towards a computational analysis of text structure

• 2010-2012

• Prof. Hajičová, Pavlína Jínová, Jiří Mírovský, Anja Nedoluzhko, Lucie Poláková, Kateřina Rysová, Magdaléna Rysová, Šárka Zikánová

• Cooperation with the Penn Discourse Treebank (Aravind Joshi, Rashmi Prasad, Bonnie Webber)

• Travelling costs (1.7 mil. CZK)
Kontakt ME10018
Towards a computational analysis of text structure

• PDTB as a resource of inspiration for the discourse annotation in the Prague Dependency Treebank
  – Consulting the methodology, comparison of the results
  – Workshops (2011 Prague, 2012 Philadelphia)

• Application for a further Kontakt project for 2013-2015
• research topics
  − past
    • valency (VALLEX, Valency lexicon of Czech verbs)
    • treebanking (grammatemes and coreference in PDT)
    • anaphora resolution, parsing, named entities
  − current
    • dependency syntax in machine translation
    • semi-supervised and unsupervised ML in NLP
• “office“
  − organizing PhD-study-related events, especially state exams and defenses
  − UFAL internships
• **Courses taught in 2012/2013:**
  
  – **Technology for NLP** (with Jan Štěpánek)
    • bash+perl+xml...
  
  – **Language data resources** (with Martin Popel)
    • corpora, treebanks, lexical databases …
  
  – **Selected Problems in Machine Learning** (with David Mareček)
    • intro to Bayesian ML, Gibbs sampling..., for PGS
  
  – **Exercises in Machine Learning** (with Ondřej Bojar)
    • gaining experience on various ML techniques

• **PhD students supervised in 2012/2013**
  
  – Nguy Giang Linh, Martin Popel, Michal Novák, Nathan Green, Loganathan Ramasamy