Short presentations

ÚFAL seminar

Sedlec-Prčice

12. – 15. 9. 2015

Last update: 2015-09-14 13:57:55, revision e582403

Contents

Petra Barančíková

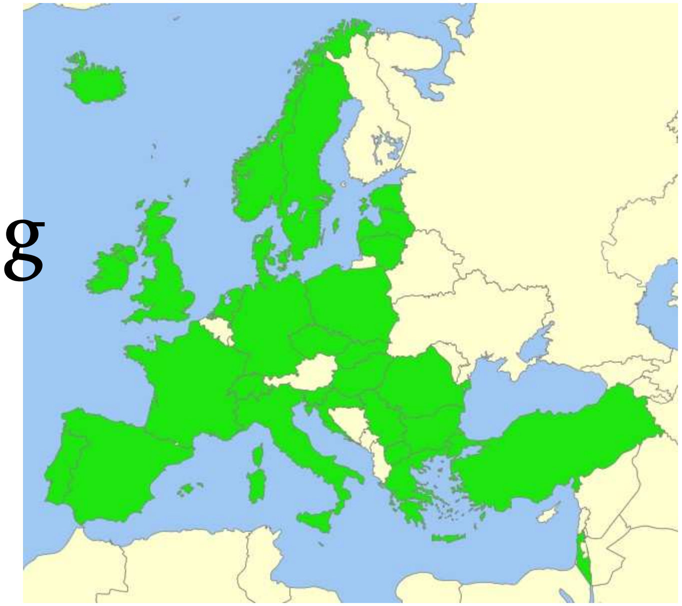
- PhD student; topic of thesis: Paraphrasing Czech Sentences for Machine Translation Evaluation (supervisor: Markéta)
- Working (¼) on Vendula's light verb GAČR
- Organizing SlonNLP workshop with Rudolf; it takes place at ITAT on September 20

Eduard Bejček

Areas of interest:

- Thesis, mainly… *yes, yes, thesis.*
- **Valency**
  - Vallink (linking of lexicons) *part of my thesis*
  - VALLEX 3.0 (data, web, book version, …)
- **Multiword expressions**
  - exploatation of manual annotation, inner structure, automatic identification *(thesis)*
- support with **LaTeX** *(e.g. PhD thesis template)*

PARSEME



- COST Action + MŠMT co-funding
- 30 european countries,

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

Current work with Pavel Straňák:

- capture the inner structure of addresses

Petra Galuščáková

- Information retrieval in audio-visual archives
  - **Retrieval** and **linking** of relevant **segments** of videos
  - Based on combination of lexical, visual and prosodic features

Hyperlinking of Video Content

- Participation in TRECVid 2015 Video Hyperlinking Task
  - Linking related segments of videos in large video archives
  - Almost 4000 hours of BBC video broadcast
- Combination of:
  - Subtitles and three automatic transcripts
  - Visual similarity based on **Feature Signatures** (in cooperation with Siret Group, KSI)
  - Visual similarity based on **Caffe descriptors** (in cooperation with DISA, MUNI)
  - **Face recognition** (in cooperation with CMP, CTU)

Anchoring in Video

- Participation in MediaEval 2015 Search and Anchoring in Video Task
- Automatic **selection of anchoring segments** in videos
  - Anchoring segments should be somehow remarkable for users of the collection.
    - Users can be interested in learning more about the topic of the anchoring segment.
  - Anchoring segments can be subsequently linked with other related segments (hyperlinking).
  - Users can browse video collection using links created for the anchoring segments.

Veronika Kolářová

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

- October 2014 – July 2015
  - Academic visit to Centre for Corpus Research (University of Birmigham)
- In Prague (UFAL):
  - LINDAT/CLARIN (30%)
- Main topics of interest
  - Valency of Czech deverbal nouns
  - Support verb constructions and their nominalizations
- Current work
  - Syntactic behaviour of nominalizations of support verb constructions
    - *giving an opportunity, loss of control*

Vincent Kríž

- PhD student (Barbora Vidová Hladká)
  - finished 3th year

- **Detecting Semantic Relations in Texts**
  - JTagger
    - detecting references in czech court decisions
    - on-line demo: ufal.mff.cuni.cz/jtagger
    - MICAI 2014, Tuxtla Gutierez, Mexico
  - RExtractor
    - detecting entities and relations in dependency trees
    - on-line demo: odcs.xrg.cz/demo-rextractor
    - MICAI 2014, Tuxtla Gutierez, Mexico
    - NAACL 2015, Denver, Colorado, USA

Vincent Kríž

- **Verb Pattern Recognition (VPR)**
  - journal paper (Cinková, Holub, Krejčová, Kríž, Materna)
  - extrinsic evaluation with MT (Bojar, Holub, Kríž)
- **Native Language Identification (NLI)**
  - experiments with language modeling (Kríž, Holub, Pecina)
  - RANLP 2015, Hissar, Bulgaria
- **Supervising**
  - 1 Bc. student

Vincent Kríž

- **Real Data Science (RDS)**
  - Barbora Vidová Hladká & Martin Nečaský
  - realdatascience.cz
  - VAVAI
    - similarities between projects → network of projects
    - suspicious places in the network
    - presentation for minister Dolezal
  - research for ČSOB
    - cooperation between MFF UK and ČSOB

Vláďa Kuboň

- Projects
  - LCT – the project still continues, the number of students is still relatively low (3 students this year)
  - GACR grant **On linguistic structure of evaluative meaning in Czech,** 2015-2017 (Katka Veselovská, Jana Šindlerová, Aleš Tamchyna)
  - participation in a GACR grant of prof.Bartak **Automatic Modeling of Knowledge and Plans for Autonomous Robots**
- Research
  - Syntactic analysis
    - Formal properties of free word order, analysis by reduction
  - MT between related languages
  - Program comittee of several workshops (EAMT, IIS, BSNLP, FLAIRS etc.)
- Teaching
  - 2 lectures - Introduction to CL and NLP Applications;
2 seminars for UFAL and
1 seminar of Automata and Grammars for CS students
  - UFAL secretary for teaching
  - Coordination of an Erasmus exchange with Saarbruecken, Koper and Tuebingen
  - Supervising 1 master thesis (defense in September),
3 PhD. students, Natálka defends next week

Jindřich Libovický (1/2)

- starting 3rd year of PhD, supervisor Pavel Pecina
- thesis topic: scene text and its recognition and hopefully soon also translation
  - cooperation with Czech Technical University (text localization)
  - signs in urban environment → specific language, many shortcuts, named entities, multiple languages
  - playing around with deep learning methods

Any one using `nolearn` and needs a visualization tool for his/her experiments?

`http://ufallab.ms.mff.cuni.cz/~libovicky/learning_curves`

Jindřich Libovický (2/2)

- maintaining collection of parallel data in medical domain (KConnect project)
  - you can find them in:

    `/net/data/medical`

  - if you have/need some, tell me
- with Ruda Rosa and Tomáš Musil, preparing a web and a short course about machine learning for high school students
  - suggestions and ideas are welcome
  - very early draft of the texts on:

    http://su.tomasm.cz/pdf/kniha.pdf

Markéta Lopatková - Projects

Research interests / research projects:

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

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

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

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

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

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

Valency lexicon:

- introduction
  - valency
  - alternations: grammaticalized and lexicalized
  - structure of the lexical entry
- grammar component
  - grammaticalized alternations – diatheses, reciprocity, reflexivity
  - lecicalized alternations – conversive, splitting, multiple
- data component

main outputs:

- printed lexicon (Karolinum) … deadline: September 30, 2015
- new on-line lexicon … deadline: December 31, 2015

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)

Central Funding

- PROVOZ (teaching money)
  - ca 2.0 mil. CZK salaries (3.65 full contracts)
  - ?? 762 th. CZK other costs
- PRVOUK (research money)
  - ca 2.8 mil. salaries (5.0 full contracts)
  - 350 th. CZK other costs
- Specific Research (?)
  - 140 th. CZK other costs

Markéta Lopatková – Teaching (1)

"Teaching projects":

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

- EM LCT (Language and Communication Technologies)
  together with Vladislav Kuboň
  3 students for 2015-16
  - 1 first year student (3 dropped)
  - 2 second year students

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

Markéta Lopatková – Teaching (2)

Courses:

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

Supervising:

- 4 PhD students, 1 Master student

Others:

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

Jiří Mírovský

Discourse-related activities

- **technical editor** of the book **Discourse and Coherence**
- **maintaining** the annotated **data**
- **searching in PML-TQ** on request
- maintaining the wiki pages with **PML-TQ examples** (mostly discourse-related)
- Management Committee and Steering Committee member of European **project COST TextLink**
- project **COST-cz TextLink** (from November)
- transforming **PDTB** data **to PML-TQ** (based on Jan Štěpánek's work)
- maintaining **TrEd extension** for **PDT 3.0** (and several others)

Jiří Mírovský

ÚFAL-wide activities

- technical support for **data checks in PDTSC** (with M. Mikulová)
- maintaining **software from LDC** (and other – dictionaries, Adobe Acrobat, ...), plus associated wiki web pages
- implementation of **analysis by reduction** on analytical trees (with M. Lopatková and V. Kuboň)
- maintaining the **Amoeba** database for ÚFAL (with V. Kuboň)
- maintaining **PML-TQ search servers** for PDT 3.0, PDiT 1.0, ...
- maintaining ÚFAL **web pages** for PDiT 1.0, PDT 3.0 (and a couple of others)
- teaching: **software project** B. Zdražilová – automatic detection of discourse relations in text
- teaching: **practical sessions** for Markéta's lectures about PDT (NPFL075)

Anja Nedoluzhko

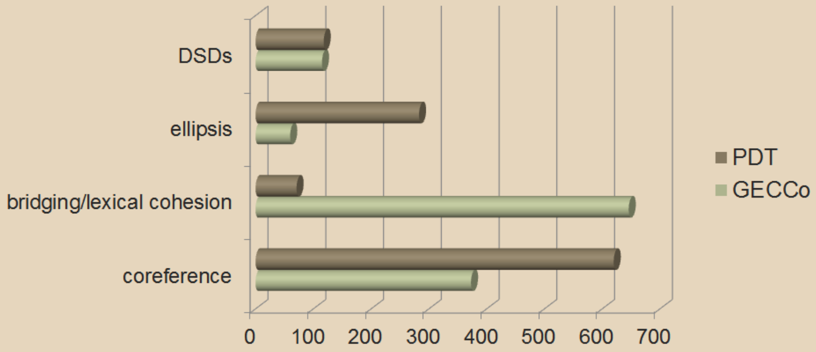
member of the **discourse team** (with prof. Hajičová, Šárka, Jirka M., Pavlína, Lucka, Bára H., Kačka and Majda Rysové, Michal Novák)

GAČR: Coreference, discourse relations and information structure in a contrastive perspective (P406/12/0658), LINDAT-CLARIN, KONTAKT (Šárka Zikánová), COST-Textlink

VÝZKUM

- prepararion of the book „Discourse and Coherence. From the Sentence Structure to Relations in Text“
  - Coreference, Bridging, Contextually bound nodes without coreference
- PDT-GECCo cooperation (next slide) – with Saarbrucken colleagues
- contextually bound nodes without coreference – with prof. Hajičová, analysis and presentation (CL)
- coreferential expressions in Czech and English (Discours) – with Michal Novák
- coreferential chains in Czech, English and Russian (Dialog) - with Michal Novák and Russian colleagues
- bridging relations multilingually – with Michal Novák and Russian colleagues
- coreference on parallel texts – with Michal Novák and foreign colleagues

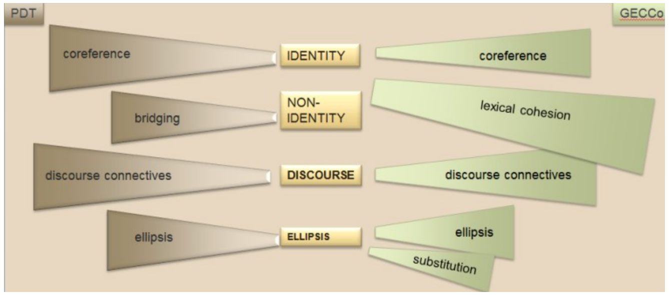
Anja: Cooperation with Saabrucken (PDT-GECCo)



Finding Nexus in the PDiT and GECCo annotation schemes (Saar, Louvain, Warszaw)

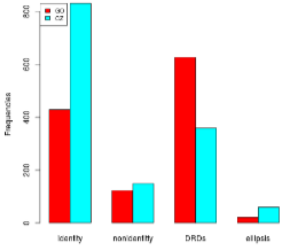
+

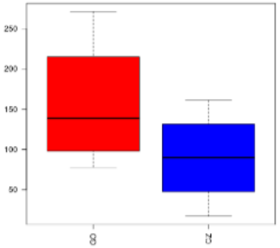
Analysis of DRD-related Contrasts in Spoken Czech, English and German (DiSpoL 2015)



Creating a Universal Annotation Scheme for Textual Relations (NAACL, LAW)

From Interoperable Annotations towards Interoperable Resources: A Multilingual Approach to the Analysis of Discourse





Anja - other

- STSM in Saarbrücken (January-February, 2015)
- oral presentations, host.prof.: Warszaw (April, 2015) and Moscow (May, 2015)
- plans for new grant proposals – with Michal Novák and discourse team (coreference resolution on parallel texts, pragmatics, cooperation with Manfred Stede about anaphoric connectives and discourse projection)
- **teaching** - Variabilities of languages (with Šárka and Magda)
- Programme Committees (LAW IX, Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-2015), International Conference on Computational Linguistics (Dialogue-2015))
- review book New perspectives on cohesion and coherence: implications for translation

Michal Novák

- GAUK 3389/2015: Cross-lingual approaches to coreference resolution
- GAČR project proposal: Structure of coreferential chains in parallel language data
  - with Anja
- coreference-related topics:
  - Anaphoric expressions and chains in parallel Czech-English (+ Russian comparable) data
    - with Anja (+ Svetlana Toldova)
  - Coreference for Deep MT into Czech and Dutch
    - with Dieke Oele, Gertjan van Noord
  - semi-supervised approaches for cross-lingual CR

Michal Novák

- QTLeap
  - 2nd year
  - TectoMT between EN and CS, NL, ES
  - adaptation for IT domain
- public service
  - supplying Karolinum bookstore with books published at ÚFAL
  - administration of the related web pages

Sales of ÚFAL books

Book	11/12 – 07/13	07/13 – 07/14	07/14 – 07/15	Total
Ondřej Bojar: Čeština a strojový překlad: Strojový překlad našincům, našinci strojovému překladu	8	1	3	12
Pavel Pecina: Lexical association measures: Collocation Extraction	3	1	2	6
Ondřej Bojar: Exploiting linguistic data in machine translation	3	1	0	4
Petr Homola: Syntatic analysis in machine translation	4	0	0	4
Anna Nědolužko: Rozšířená textová koreference a asociační anafora	3	1	0	4
Barbora Štěpánková: Aktualizátory ve výstavbě textu, zejména z pohledu aktuálního členění	-	-	4	4
Marie Mikulová: Významová reprezentace elipsy	2	1	0	3
Jiří Mírovský: Searching in the Prague Dependency Treebank	1	2	0	3
Zdeňka Urešová: Valence sloves v Pražském závislostním korpusu	3	0	0	3
Zdeňka Urešová: Valenční slovník Pražského závislostního korpusu PDT-Vallex	3	0	0	3
Kateřina Rysová: O slovosledu z komunikačního pohledu	-	-	3	3
Magda Ševčíková: Funkce kondicionálu z hlediska významové roviny	1	1	0	2
Silvie Cinková: Words that Matter: Towards a Swedish-Czech Colligational Dictionary of Basic Verbs	1	0	0	1
Total	32	8	12	**40**
Book / Month	4.0	0.67	1.0	**1.25**

Donations of ÚFAL books

Book	#
Ondřej Bojar: Exploiting linguistic data in machine translation	18
Pavel Pecina: Lexical association measures: Collocation Extraction	17
Petr Homola: Syntatic analysis in machine translation	15
Jiří Mírovský: Searching in the Prague Dependency Treebank	4
Silvie Cinková: Words that Matter: Towards a Swedish-Czech Colligational Dictionary of Basic Verbs	4
Ondřej Bojar: Čeština a strojový překlad: Strojový překlad našincům, našinci strojovému překladu	2
Marie Mikulová: Významová reprezentace elipsy	1
Total	61

- donated during events organized by UFAL:
  - SIGDIAL 2015
  - Deep MT Workshop 2015
  - MT Marathon 2015
  - Prague-Hamburg seminar 2015

Lucie Poláková

Project:

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

Recently:

- PDT 3.0 released in December 2013 (new: mainly genres, rhematizers, Anja – coreference of 1st + 2nd person)
- Annotation of alternative lexicalizations od discourse connective in PDT (Majda Rysová)
- Preliminary research on implicit connectives (no discourse connective present) and on attribution ("who said what")

Grant support:

GAČR (Zikánová, "Interplays" till 2015)

LINDAT

KONTAKT, second round - Cooperation with UPenn and prof. Aravind Joshi's team (August 2015 – trip to dr. R. Prasad in Milwaukee, Winconsin – **PennDTB 3.0 soon to appear!)**

Lucie Poláková

COST: TextLink (Action IS1312)

Huge EU project on multilingual corpora with linked connectives – funding mainly for gathering scientists and resources, newly: a tiny amount for salaries (PI – L. Degand, Belgium)

October 20.-21. 2014 meeting in Prague (L. Degand, B. Webber, M. Stede etc.); April 2015 – Fribourg Switzerland (T. Sanders, S. Zufferey) – unified discourse-semantic classification for European languages

PhD: Dissertation:

Discourse relations in Czech, defence on September 23rd, 2015

Service:

- new UFAL webpages
- nástěnkářka ☺

Martin Popel

- **Treex** NLP framework (Treex::Web no progress since last year)
  - SVN → GitHub (https://github.com/ufal/treex)

    pull requests, issues, forks, Travis-CI, smaller codebase
  - Perl scenarios: easy to use, versioned, parameters

    ```
    treex -Len -Ssrc Read::Sentences from=en.txt
    Scen::EN2CS Write::Sentences to=cs.txt
    ```
- **TectoMT** machine translation (PhD thesis on transfer)
  - now for EN↔CS, EN↔ES, EN↔NL, EN↔PT, EN↔EU
  - experiments with Vowpal Wabbit and word embeddings
- **HamleDT 3.0** (42 treebanks), Dan's GAČR **Manyla**
- **MT-ComparEval** (with Ondřej Klejch, paper in PBML/MTM)

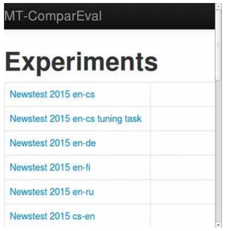
  https://github.com/choko/MT-ComparEval

MT-ComparEval:
Graphical evaluation interface for Machine Translation development

Ondřej Klejch, Eleftherios Avramidis, Aljoscha Burchardt, Martin Popel
klejch@ufal.mff.cuni.cz, {eleftherios.avramidis,aljoscha.burchardt}@dfki.de, popel@ufal.mff.cuni.cz
Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
German Research Center for Artificial Intelligence (DFKI), Language Technology Lab

- **Try it now (all WMT 2014–2015 results):**
  http://wmt.ufal.cz
- **Install it** (and report issues or contribute):
  https://github.com/choko/MT-ComparEval

- **web-based tool for MT developers**
- **check progress of a system over time or compare several MT systems**
- **focus on analyzing system differences**
- **can be integrated into your workflow (import translations from disk/git/...)**

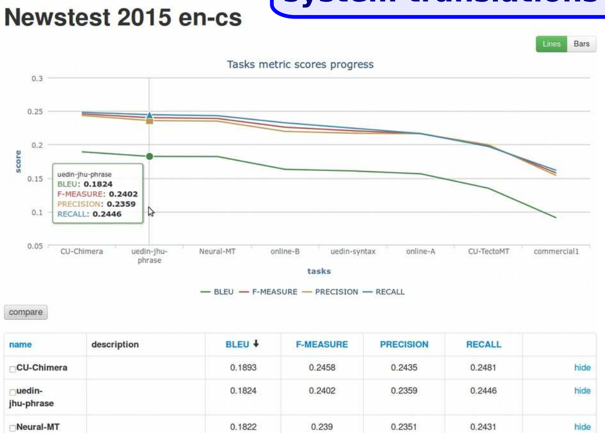

MT-ComparEval

Experiments

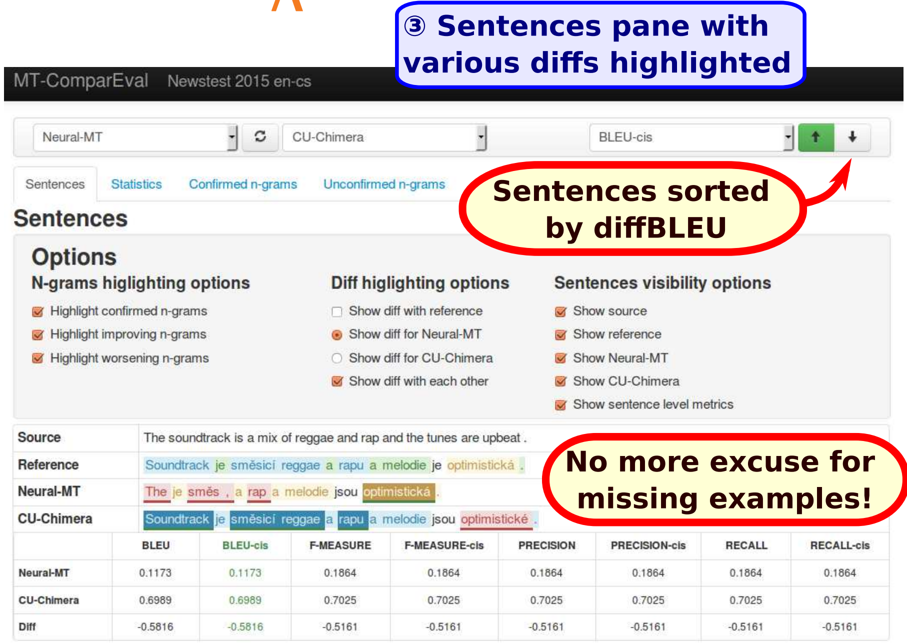
- Newstest 2015 en-cs
- Newstest 2015 en-cs tuning task
- Newstest 2015 en-de
- Newstest 2015 en-fi
- Newstest 2015 en-ru
- Newstest 2015 cs-en

① Select an "Experiment", e.g. English-Czech WMT15

② See BLEU, F-measure,... Select two "Tasks", i.e. system translations


Newstest 2015 en-cs

name	description	BLEU ↓	F-MEASURE	PRECISION	RECALL	
□CU-Chimera		0.1893	0.2458	0.2435	0.2481	hide
□uedin-jhu-phrase		0.1824	0.2402	0.2359	0.2446	hide
□Neural-MT		0.1822	0.239	0.2351	0.2431	hide


③ Sentences pane with various diffs highlighted

MT-ComparEval Newstest 2015 en-cs

Sentences sorted by diffBLEU

Sentences

Options

N-grams higlighting options
- ☑ Highlight confirmed n-grams
- ☑ Highlight improving n-grams
- ☑ Highlight worsening n-grams

Diff higlighting options
- □ Show diff with reference
- ● Show diff for Neural-MT
- ○ Show diff for CU-Chimera
- ☑ Show diff with each other

Sentences visibility options
- ☑ Show source
- ☑ Show reference
- ☑ Show Neural-MT
- ☑ Show CU-Chimera
- ☑ Show sentence level metrics

Source	The soundtrack is a mix of reggae and rap and the tunes are upbeat .
Reference	Soundtrack je směsicí reggae a rapu a melodie je optimistická .
Neural-MT	The je směs , a rap a melodie jsou optimistická .
CU-Chimera	Soundtrack je směsicí reggae a rapu a melodie jsou optimistické .

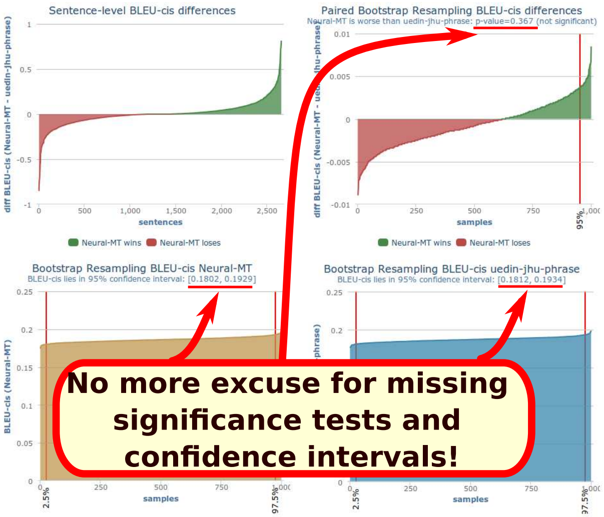
	BLEU	BLEU-cis	F-MEASURE	F-MEASURE-cis	PRECISION	PRECISION-cis	RECALL	RECALL-cis
Neural-MT	0.1173	0.1173	0.1864	0.1864	0.1864	0.1864	0.1864	0.1864
CU-Chimera	0.6989	0.6989	0.7025	0.7025	0.7025	0.7025	0.7025	0.7025
Diff	-0.5816	-0.5816	-0.5161	-0.5161	-0.5161	-0.5161	-0.5161	-0.5161

No more excuse for missing examples!

- **confirmed n-gram** = occurres in the reference (light yellow and blue highlight)
- **improving n-gram** = confirmed n-gram occurring in only one of the systems (dark yellow and blue highlight)
- **worsening n-gram** = unconfirmed, occuring in only one of the systems (red highlighting)
- **diff** (LCS underlined in green)

- See the sentences with biggest improvement/worsening
- See the 1-grams...4-grams with biggest improvement/worsening
- Hints for improving the systems

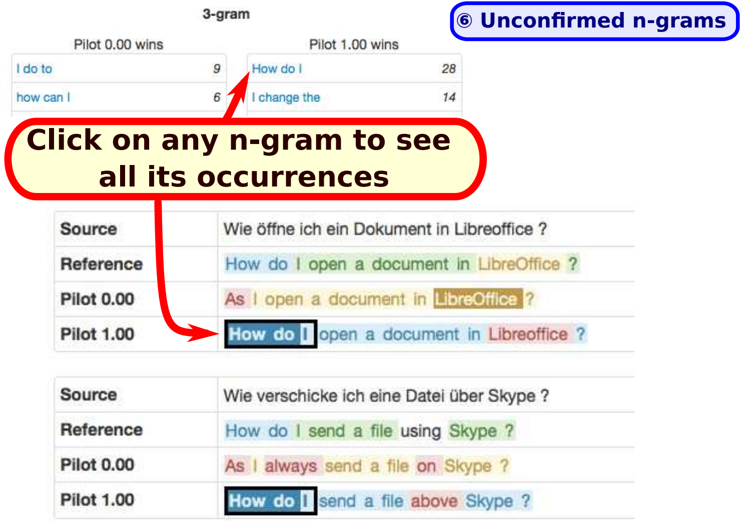
④ Statistics pane (paired bootstrap resampling,...)



No more excuse for missing significance tests and confidence intervals!

⑤ Confirmed n-grams pane

⑥ Unconfirmed n-grams


3-gram

Pilot 0.00 wins		Pilot 1.00 wins	
I do to	9	How do I	28
how can I	6	I change the	14

Click on any n-gram to see all its occurrences

Source	Wie öffne ich ein Dokument in Libreoffice ?
Reference	How do I open a document in LibreOffice ?
Pilot 0.00	As I open a document in LibreOffice ?
Pilot 1.00	How do I open a document in Libreoffice ?

Source	Wie verschicke ich eine Datei über Skype ?
Reference	How do I send a file using Skype ?
Pilot 0.00	As I always send a file on Skype ?
Pilot 1.00	How do I send a file above Skype ?

Presented at MT-Marathon 2015, Prague, September. Supported by the FP7 EU grant QTLeap (No. 610516), and SVV project 260 104. Using language resources hosted by the LINDAT/CLARIN Research Infrastructure, Project No. LM2010013 of the Czech Ministry of Education, Youth and Sports.
Published in The Prague Bulletin of Mathematical Linguistics, No. 104, pp. 63–74. https://ufal.mff.cuni.cz/pbml/104/art-klejch-et-al.pdf doi: 10.1515/pralin-2015-0014

Martin Popel

- **PBML** (next deadline: January 15th 2016) LaTeX volunteer?
- **Technical reports** (2015 deadline: December 1st)
- **Teaching** autumn: Modern Methods in CL I ("Reading group")

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

  Language Data Resources (with ZŽ, me: significance)
- **QTLeap** FP7 project, 2013–2016
  (14 lang. pairs, 8 partners: FCUL, DFKI, CUNI, IICT-BAS, UBER, UPV/EHU, UG, HF)

  TectoMT outperforms Pilot0 (Moses) on the IT domain
  (QTLeapCorpus) for: EN↔CS, EN→PT, EN→ES, NL→EN

  TectoMT available via MT-Monkey web service
  and (more) easily installable

Rudolf Rosa

- semi-supervised cross-lingual syntactic analysis
  - PhD (ZŽ, starting 3rd year); GAUK; IWPT, ACL, Depling
- QTLeap (TectoMT), HiML (Depfix/MLfix)
- tecto paraphrasing for MT eval (helping Petra B.)
- workshops: SloNLP organizer, DMTW local support
- HamleDT, Universal Dependencies (mildly involved)
- TA for ZŽ's Technology for NLP (last year)
- interning at Google Zürich until March 2016
  - some NLP research (confidential), with Srini Narayanan

Magda Ševčíková

involved in projects

- LM2010013 LINDAT-Clarin

teaching

- *Introduction to Formal Linguistics*
  - for master students, Faculty of Mathematics and Physics
- *Professional language and style*
  - course on academic writing
  - for master students, Faculty of Mathematics and Physics
- *Variability of languages in time and space*
  - with Anja Nedoluzhko and Šárka Zikánová
  - for PhD students, Faculty of Mathematics and Physics
- *Modern linguistic descriptions of English*
  - course on selected syntactic theories
  - for master students of English philology, Faculty of Arts

DeriNet
Lexical network of derivational relations in Czech (i)

- lemmas (nodes) connected with links (edges) corresponding to derivational relations
- with Zdeněk Žabokrtský, Jonáš Vidra (Bc. thesis; SFG grants), Adéla Limburská (SFG grant), and Milan Straka (DeriNet Viewer, expert's opinion on inflectional data)
- Jonáš's thesis
  - extending the DeriNet 0.9 data to MorfFLex dictionary ($>$ DeriNet 1.0)
- current topics
  - representing lemmas with homonymy indexes from MorfFlex
  - exploiting alternations in inflectional paradigms for searching for derivational pairs
    - inflection: *d**ů**m*.nom – *d**o**mu*.gen, *sn**í**h*.nom – *sn**ě**hu*.gen
    - derivation: *d**ů**m* – *d**o**mek*, *sn**í**h* – *sn**ě**žný*
  - semantic labelling of derivational relations
    - high ambiguity vs. homonymy of derivational affixes in Czech

DeriNet
Lexical network of derivational relations in Czech (ii)

- `http://ufal.mff.cuni.cz/derinet`
  - data of DeriNet 0.9
- version 1.0 to be released soon in LINDAT/Clarin repository
- *DeriNet Viewer* tool by Milan Straka
  - `http://ufal.mff.cuni.cz/derinet/viewer`
  - DeriNet 0.9
- *DeriNet Search* tool by Jonáš Vidra
  - `http://jonys.cz/derinet/search/`
  - DeriNet 1.0

Aleš Tamchyna

- Ph.D. student, finishing the third year.
- Advisor: RNDr. Ondřej Bojar, Ph.D.
- Research interests:
  - statistical machine translation,
  - machine learning in natural language processing.
- Thesis topic: Lexical and Morphological Choices in Machine Translation.
- I will spend the next academic year in Munich.

Zdeňka Urešová

1. GAČR POSTDOC PROJECT (2013 – 2015)

A comparison of Czech and English verbal valency based on corpus material

- Data preparation together with Jana Šindlerová, Eva Fučíková
- Related: AMR annotation comparison between Czech and English
  - Collaboration with Martha Palmer, Claire Bonial, Jena D. Hwang and Nianwen Xue
- Publ. in 2015 – all 5 related to CzEngVallex (The 9th Linguistic Annotation Workshop 2015, co-located with NAACL 2015, SemEval 2015, Depling 2015 2x, Technical report)

CzEngVallex
(both sides of the PCEDT)
20,835 aligned frame pairs
including argument alignment
http://lindat.mff.cuni.cz/services/... coming soon!

2. AMALACH PROJECT

- Translation checking and translation interface testing

3. Collaboration with Adam Przepiórkowski

- PDT-Vallex vs. Walenty paper

New submitted proposals for next years:

- GAČR: Verbal meaning relations based on an annotated parallel corpus acceptance still pending
- PIRE: eCL-AMR: Exploring Cross-linguistic Abstract Meaning Representations not accepted ☹
- KONTAKT: Abstraktní reprezentace významu a zpracování přirozeného jazyka not accepted ☹

Kateřina Veselovská

- about to finish a Ph.D.

  „On the Linguistic Structure of Emotional Meaning in Czech“

- in fact

  :(SEANCe:) = SEntiment ANalysis in Czech

- sentiment-annotated corpus SubLex
- manually annotated data from multiple domains
- several polarity classifiers with rather satisfactory results (89% accuracy)
- implementation of SubLex to PDT
- new GUI in TrEd
- annotation guidelines (technical report, in progress)
- opinion-target detection systems (CZ/EN)

Other ‘sentimental’ projects:

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

:(Kateřina Veselovská:)

Other topics of interest:

- construction grammar
- tectogrammatical description of English
- multimodal corpora
- teaching: Linguistic Applications (FF UK, FF UPOL)
- theses consultations & supervisions
- text analytics, i.e. business applications of text mining

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

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

Dan Zeman

- Funding
  - MANYLA (GAČR)
  - HIML => translation of health brochures, morphology-aware, Polish and Romanian
  - PRVOUK
  - LINDAT

Dan Zeman

- Funding
  - MANYLA (GAČR)
  - HIML => translation of health brochures, morphology-aware, Polish and Romanian
  - PRVOUK
  - LINDAT

Dan Zeman

- **Interset:** conversion of morphosyntactic tags between tagsets (even cross-language)
  - Universal description of morphological tagsets
- **HamleDT**
  many treebanks → common annotation style
  - With Martin P., Zdeněk, Ruda and others
- **Universal Dependencies**
  - Large community effort

Dan Zeman

- Parsing
  - Under-resourced languages, delexicalized parsing
  - Maltese treebank
- Teaching
  - Morphological and Syntactic Analysis I & II
  - Disrupted: Computational NLP (but kept at ČVUT)
- Other
  - Bibliography maintenance (the “Biblio” database)

Šárka Zikánová

Projects

- GA ČR: standard project

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

- KONTAKT: cooperation with University of Pennsylvania

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

Tasks

- Main coordinator of two representative publications of the discourse team
  - Slovo a slovesnost
  - monograph Discourse and Coherence: From the Sentence Structure to Relations in Text
- Fund raising for the discourse team

Teaching

Courses:

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

- Supervisor of Ph.D. theses and master theses