LINDAT/CLARIN
Language Research Infrastructure

Jan Hajič
principal investigator, LINDAT/CLARIN
Workshop on Digital Humanities
Prague, Sep. 24, 2015
LINDAT/CLARIN

• Infrastructure for Language Research in...
  – Humanities (linguistics and related)
  – Computational Linguistics

• Provides
  – Data repository, linked to European networks
    • Clarin ERIC, META-SHARE
  – National Center for language data
    • Acquisition, specification, licensing
    • Linguistic and language-related annotation
    • Archiving, preservation
    • Distribution
LINDAT/CLARIN Partners

- Charles University in Prague (coordinator)
  - Institute of Formal and Applied Linguistics
    - Faculty of Mathematics and Physics, CS School
- Masaryk University (Brno)
  - Natural Language Lab
    - Faculty of Informatics
- University of West Bohemia (Pilsen)
  - Dept. of Cybernetics (Speech and Language group)
    - Faculty of Applied Sciences
- Czech Language Institute
  - Academy of Sciences of the Czech Rep.
LINDAT/CLARIN

• Project Timeframe
  – 2010-2015: funding secured (EUR 4.7M)
    • Started in October 2010
    • 2010: building the technical backbone, hiring
      – New server room
      – Ready for expansion (up to 400 CPU, 100TB)
  • 2011-2013
    – Construction phase: node of Clarin ERIC (Feb. 2012)
      » Compatibility with other networks (META-SHARE)
  • 2014-
    – “Operational” phase
Welcome to LINDAT/CLARIN

Centre for Language Research Infrastructure in the Czech Republic

The LINDAT/CLARIN Centre for Language Research Infrastructure provides technical background and assistance to institutions or researchers who want to share, create and modernise their tools and data used for research in linguistics or related research fields. The project also provides an open digital repository and archive open to all academics who want their work to be preserved, promoted and made widely available. 

LINDAT/CLARIN is funded by the Ministry of Education, Youth and Sports of the Czech Republic.
LINDAT/CLARIN Project structure

• Clarin ERIC coordination
  – National Coordinator
• Clarin “Centre” type B (since Jan. 2014)
  – Node of the Clarin ERIC network
  – 24/7 service, data repository, full compatibility
  – Software and services development, contracts and agreements with service providers, licensing
• National Centre for Czech Language Resources
  • acquisition, conversion, description (metadata)
  • annotation, preservation
  • evaluation data, shared tasks preparation, technology services
  • methodology development for resource creation
Data-related services
- Repository (dSpace, v5.3)
  - Universal access and authentication
    - Single sign-on (Shibboleth/EduGain, local) within Clarin ERIC
    - Permissions, licensing (active, passive) – built-in license selector
  - Persistency, archiving, metadata API (Clarin ERIC, META)
    - Persistent ID assignment (EPIC handles – our own handle server)

Language services
- Clarin ERIC compatible API (final goal: Weblicht-compatibility)
- Workflows (treex now fully integrated)
- (Mostly) Czech-language related
  - morphology, tagging, parsing, anaphora resolution, …
- Services running now, integrated (14, more to come)
LINDAT/CLARIN
Building Czech Resources

• All partners… history of resource creation
  – Charles University in Prague
    • Prague Dependency Treebank, parallel corpora for Machine Translation, spoken language corpora, NLP tools
    • Digital Libraries: The Malach project (Visual History, USC)
  – University of West Bohemia
    • Speech-related data (ASR, TTS), speech tools
  – Masaryk University
    • Lexical resources and tools, Federated content search
  – Institute of Czech Language
    • Lexical and diachronic language resources, Language Guide

• LINDAT/CLARIN: continued resource building
LINDAT/CLARIN
Education and Training

• University partners
  – Master’s and Ph.D. programmes
    • Linguistics, Computer Science, NLP, Machine Learning…
    • New accreditation: two Master’s programs, CL and ML/NLP

• LINDAT opportunities for students
  – NLP and speech-related experiments
  – new data specification annotation
  – linguistic research

• Training
  – Czech researchers & international
    • January 2015: 50 participants, PARSEME (COST EU action)
    • Summer 2014: Johns Hopkins University Engineering workshop, 55 part.
LINDAT/CLARIN
International Cooperation

• Language Resources (world-wide)
  – Same needs, similar datasets
  – Since early 1990s (Statistical methods in NLP)

• History of cooperation
  – LDC (since 2001), ELDA (since 2005)
    • MU, UWB and CUNI has already provided some datasets
      – PDT, U.S. Malach project, Czech Wordnet, parallel corpora
      – Evaluation datasets (CLEF, CoNLL, MT – several years)

• Cooperation with research institutions
  – Assessment of needs, research plans
  – Future / emerging technologies: data first, “Big data”
  – Tools (latest results – high quality tools, including for Czech)
LINDAT/CLARIN

• Three areas
  – Building the repository, standards, compatibility
  – Filling in the repository – Czech, other language data (annotation)
  – Coordination inside/outside

• Current status and statistics:
  – LINDAT/CLARIN operational since Jan. 1, 2014
    • Certified Clarin B Centre, (incl. DSA), working on A-class certification
  – 14 services / demos integrated, plus the ILG (ÚJČ / A.S.)
  – More data in repository, incl. part of CNC (134 items)
    • 5 billion words, searchable by KonText (cooperation with UCNK)
  – Language Resources and Tools, from Clarin: 973 items
  – 30,000 accesses/months, lexical resources: e.g. 55,000 views PDT-Vallex
  – Internet Language Guide: 9 mil. accesses/year
Providing expertise and resources to outside projects

• First example, continued from 2012-2013
  – Project for annotated corpus of foreign learners of Czech
    • same for „disadvantaged“ minorities (= Roma/Gypsy)
  – Cooperation with FF UK Prague
    • prof. Karel Šebesta and many others, ÚTKL FF UK

• Providing support, repository
  – Expertise with data preparation, metadata
  – Small amount for support of annotation

• 1 additional corpus in 2014 (AKCES 5 CzeSL- SGT)
  – “Czech as a Second Language with Spelling, Grammar and Tags”
    • AKCES 2 versions 1 and 2 from 2013, language learning, anonymized
    • AKCES 3 and 4 since 2012 (learners of Czech, disadvantaged minorities)
Tools and Services I

- LINDAT lexical resources (since end of 2011)
  - PDT-Vallex 2.5+PDT: [http://lindat.mff.cuni.cz/services/PDT-Vallex](http://lindat.mff.cuni.cz/services/PDT-Vallex)
  - CzEngVallex: [http://lindat.mff.cuni.cz/services/CzEngVallex](http://lindat.mff.cuni.cz/services/CzEngVallex) (coming soon)
- PML-TQ – public access & search for treebanks
  - 40 now public, other in negotiations
- Treex::web – service-based NLP tools, integrated
- Morphology analysis and generation (Czech), MorphoDiTa
  - New version – universal morphology, tagger, generator
- Sample treebank & spoken corpora
  - simple graphical browsing (e.g. [http://ufal.mff.cuni.cz/pcedt2.0](http://ufal.mff.cuni.cz/pcedt2.0))

Sep. 24, 2015

LINDAT/CLARIN
Tools and services II

• Tools developed outside of LINDAT/CLARIN

• Machine translation
  – Translation Czech ↔ English
    • (‘14 versions of Moses, TectoMT/Depfix)
  – Translation Czech → Slovak (updated, integrated)
  – Other languages

• Re-accentization, spelling correction
  – http://lindat.mff.cuni.cz/services/korektor

• Tutorials (tred, PML-TQ)
  – Annotation tool tred, treebank search tool PML-TQ
Other activities and figures

- **Publications in 2014**
  - 5 books, 5 chapters in books, 28 journal articles, 64 proceedings papers (total: 98)
- **Licenses “granted”**
  - 39101 (2013: 7147) downloads, website views: 154,588; 11 mil. IJP
- **Total FTEs**: 21.35, all 4 sites, total
- **21 annotators finishing annotations, transcriptions (~ 6 FTE)**
  - Only those with “insured” temporary contracts; more on short-term contracts
  - 7 teams for annotations (2013: 10)
- **Students working on repository, data, tools, web: ~5**
  - No. of Ph.D. graduates in 2014 using data from LINDAT/CLARIN: 5
- **Conferences (co-)organized: 5**
  - Coling 2014, JHU Workshop 2014, TSD by MU Brno/UWB Pilsen, WMT, MTM
- **International cooperation**: 8 projects using/delivering data
  - 2015: Four H2020 projects starting using LINDAT/CLARIN for data (and/or repo)
  - Three proposals submitted in H2020 (Eudat2, INFRADEV-3, VRE/INTERVIEWS)
- **Membership in boards of infrastructures:**
  - SCCTC Clarin, IAP Clarin NL, TAB Clarin/Dariah DE
http://lindat.cz

Thank you for your attention!