LINDAT-Clarin
Language Research Infrastructure

Jan Hajič

Charles University in Prague

Czech Republic

LINDAT Project principal investigator

LM2010013
Supported by

Research in the area of Natural Language

- Language research paradigm:
  - Empirical research
    - Theoretical linguistics: evidence, counterexamples, …
    - Quantitative and formal linguistics: preferences, patterns, …
    - Computational linguistics / Natural Language Processing
      - Statistical (machine) learning methods prevalent
- Need for language-related data
  - Language Corpora
    - Raw & analyzed (manually, automatically)
    - Monolingual, parallel
    - Text, audio, video, multimedia
  - As much as possible, easy (free) access

Research Infrastructure

- Provides support for research
  - Measurements, data, expertise, …
- Language infrastructure
  - Language and language-related data
  - Easy access and licensing
  - Possibility to include researcher's data
  - Tools for language processing
    - Web services

LM2010013
Supported by

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
  - Part of the 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

LM2010013
Supported by

LINDAT-Clarin Project structure

- Clarin ERIC coordination (Netherlands, MPI)
  - National Coordinator
- Full Clarin "Centre"
  - Node of the Clarin ERIC network
  - 24/24 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
    - methodology development (e.g. crowdsourcing)

LINDAT-Clarin
Clarin Centre

- Data-related services
  - Repository (dSpace)
    - Universal access and authentication
      - Single sign-on (Shibboleth) within Clarin ERIC
        - » 2011: problematic in several respects; local registration added
      - Permissions, licensing (active, passive)
    - Persistency, archiving, metadata API (Clarin ERIC, META)
      - Persistent ID assignment (EPIC handles)
- Language services (plans only, not started yet)
  - Clarin ERIC compatible API (Weblicht?), workflows
  - (Mostly) Czech-language related
    - morphology, tagging, parsing, anaphora resolution, …

LM2010013
Supported by

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
  - Institute of Czech Language
    - Lexical and historic language resources
- LINDAT-Clarin: continued resource building

LINDAT-Clarin Education and Training

- University partners
  - Master's and Ph.D. programmes
    - Linguistics, Computer Science, NLP, …
- LINDAT opportunities for students
  - NLP and speech-related experiments
  - new data specification annotation
  - linguistic research
- Training
  - Winter school(s), tutorials for researchers
  - Czech researchers & international

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 organizations
  - Assessment of needs, research plans
  - Future / emerging technologies: data first
  - Tools (latest results – high quality tools, including for Czech)

Access to Resources

- Repository
  - Technological use – clear (download)
    - Accompanied with software fro technological use
  - Problem: use in humanities, some licenses
- Use in humanities (Clarin-like)
  - Plan: web applications based on web services
  - Current technology:
    - Web-based access
    - Example:
      - LINDAT lexical resources
        - » VALLEX 2.6: http://ufal.mff.cuni.cz/vallex/2.6
        - » PDT-Vallex 2.5+PDT: http://ufal.mff.cuni.cz/lindat/PDT-Vallex.html
      - Dialogs: http://ujc.dialogy.cz, http://ufal.mff.cuni.cz/~toman/pdtsc, http://ufal.mff.cuni.cz/~toman/pdtse
      - Treebanks
        - » PEDT http://ufal.mff.cuni.cz/pedt2.0
        - » PCEDT http://ufal.mff.cuni.cz/pcedt2.0

Plans for 2012 and beyond: repository and coordination

- Clarin ERIC – finally!
  - Start collaboration with Clarin-D, NL etc.
  - Authentication – top priority
    - Present licensing solution developed at UFAL
  - Central site? (search / metadata sharing)
    - Learn from META-SHARE experience
  - Implementation of Clarin standards
  - Compatibility (conversions) with META-SHARE
  - Legal issues (cooperation with Clarin-D)

Plans for 2012 – Czech-related data

- Acquisition of data
  - Goal:
    - 2 GW in repository
    - 5 GW total (searchable)
- Data annotation
  - Towards PDT 3.0:
    - Bridging anaphora
    - Discourse
    - Corrections
  - Spoken annotation
    - PDT-like annotation on (reconstructed) spoken corpora

LM2010013
Supported by

Plans for 2012 - dissemination

- LINDAT-Clarin
  - Lectures at various events
    - CLARA EU winter school – Feb. (tutorial, treebanks)
    - Workshop at LREC (META-NET organized)
      - Advanced treebank annotation
- Offering services
  - Web-based, access to large corpora
  - Annotation services
    - such as the tasks done for ELRA on small scale in 2010/11

LINDAT-Clarin

http://www.lindat.cz

Thank you for your attention!

LM2010013
Supported by