CLARIN-PL – A Minute Before

Maciej Piasecki, Bartosz Broda
Wrocław University of Technology
G4.19 Research Group
maciej.piasecki@pwr.wroc.pl

CLARIN Conference
Sofia
2012-10-25-27

Infrastructure

- **CLARIN-PL Language Technology Centre**
  - **B-type centre**, located in Wrocław University of Technology
  - repository for **resources** from the partners and beyond
  - Web Services implementing **tools** for the basic processing chain of Polish
  - **web applications** developed for the selected H&SS partners
  - also an active **K-type centre** in several areas

    (more K-centres possible during the Exploitation Phase)
- Partners:
  - 3 Language Technology partners,
  - 1 corpus technology and linguistics partner,
  - 2 corpus linguistic partners

Language Technology Centre

- Implementation of all requirements is planned, e.g.:
  - a proper repository system (supporting persistent identifiers for resources and tools)
  - application of all CLARIN standards and protocols
  - participation in the national identity federations and in the CLARIN service provider federation
  - support for meta-data harvesting and distributed, co-operation with supercomputing centres
  - opportunities for researchers from H&SS to integrate the offered resources into their workflows
  - coordination of CLARIN-PL on the technical level
- Construction
  - to be started in 2013 and completed by 2015
  - based on open licence software and existing CLARIN services

Prototype and Web Services

- Prototype **http://nlp.pwr.wroc.pl/clarin**
  - Morpho-syntactic tagging (***TaKIPI-WS***)
  - Access to corpus-based similarity between words (***SuperMatrix-WS***)
  - Access to plWordNet (***plWordNet-WS***)
- Other web services not yet included in the infrastructure (completed or in construction, SyNaT project)
  - Morphological analysis (*Morpho-WS*)
  - Morpho-syntactic tagging (*Tagger-WS* based on *WCRFT* tagger)
  - Chunking and recognition of syntactic relations between chunks (*Chunker-WS*, *IOBBER-WS*, *ChunkRel-WS*)
  - Named Entity recognition (***NER-WS***) – completed
  - Recognition of semantic relations between named entities (*Serel-WS*)
  - Anaphora resolution (*IKAR-WS*)
  - Word Sense Disambiguation (*WSD-WS*)

Tasks for CLARIN-PL

- CLARIN-PL Language Technology Centre construction
- A system for long term preservation of digital data
- Corpora: speech, transcription, historical and bilingual
- Tools for advanced searching text and speech corpora and linguistic knowledge extraction
- Lexical semantics resources: huge bilingual wordnets, multi-word expressions, Proper Names and valency frames
- Shallow and deep semantic parsers for Polish
- Information Extraction: Named Entities, relations, situations
- Text Summarisation tool
- Text Mining focused on applications in H&SS – cooperation with H&SS users

Thank you for your attention