Language Technology

underlying technologies for research infrastructure in the social sciences and humanities

Tamás Váradi
Hungarian Academy of Sciences
varadi@nytud.hu

The Role of Language

- Dominant medium now and in future
- Congenial for humans, challenge for machines
- not merely streams of characters, words but rather concepts, properties, relations, arguments, facts etc.
- accessible to humans, partially so to machines
- information explosion of electronic texts
- tapping the huge resources of the web

Language Technology

- Intelligent handling of language through automated procedures
  - information extraction/data mining
  - text understanding
  - speech recognition
  - semantic web
  - multilingual information extraction
  - machine translation

Language Resources

- Data
  - Corpora – linguistically annotated textual databases
    - monolingual and parallel aligned corpora
  - lexicons
  - terminological databases
  - ontologies
- Annotation technology
- Tools

Tools

- Tokenizers
- Morphological analyzers
- Taggers
- Syntactic parsers
- Semantic taggers
- Parallel corpus aligners
- Automatic term extractors
- Intelligent data mining

The TELRI Story

- 1996 – 2002 EU founded networking project
- A truly Pan-European exercise
- Consortium of 26 nations involving 24 languages
- Tried and tested network of research centres of excellences
- Representing state of the art language technology expertise

TRACTOR

TELRI RESEARCH ARCHIVE OF COMPUTATIONAL TOOLS AND RESOURCES

- Massive collection of language resources in 24 languages
- Accessible on-line and available free for research
- Currently hosted by Oxford University
- Not (just) for linguists and not necessarily by linguists!

Summary

- Modern research methodology in SSH requires effective means to deal with the bottleneck presented by the explosive growth of textual data in electronic form
- Language technology is a fundamental enabling infrastructure that can be deployed in many areas to address the need to process texts intelligently
- SSH research infrastructure should greatly benefit from the utilization of state of the art language technologies