Building the Croatian Language Technology Portal

Marko Tadić, Ivana Simeon
(marko.tadic@ffzg.hr, ivana.simeon@ffzg.hr)
Department of linguistics
Institute of linguistics
Faculty of philosophy
University of Zagreb (www.ffzg.hr)

CUC2001, Zagreb, 2000-09-25

Croatian Language on Web

- data on Croatian language
  - dispersed across different sites
  - institutionally and geographically heterogeneous
  - professional or amateur — hard to tell
- no professional, systematical, institutionally supported homepage for Croatian Language
  - Institute for Croatian Language and Linguistics?
  - Dept. of Croatistics, FFZG?
  - Ministry of Science and Technology?

Croatian Language on Web 2

- the web-presence of national language
  - not just the number of pages published in that language
  - existence and on-line accessibility of basic language resources
    - representative (national) and specialized corpora
    - general and specialized dictionaries
    - possible MT service (like AltaVista with Systran engine)
- web-presence impossible without developing Language Technologies for that language
  - Croatian?
  - CUC2000!

Language Technologies for Croatian

- attempt to take a snapshot of situation
- *Croatian Language Technologies Portal*
  - i-project
  - started November 2000
  - funded by MZT RH under grant No. 00-86
  - duration: 1 year
  - maintaining data: 3 more years at least

Human Language Technologies

- defined in EU Framework Programme 5
  - continues also within FP6
- main research area:
  - IST = Information Society Technologies (26.3% budget of FP5 or 3,600 M€)
- key action III of IST:
  - MC&T = Multimedia Content and Tools (564 M€)
- largest part of MC&T:
  - Human Language Technologies = HLT
- interdisciplinary field: linguistics & IT

HLT 2

- language resources
  - corpora
  - dictionaries
- language tools
  - morphology
    - generators/analyzers
    - POS taggers
  - syntax
    - shallow/deep/robust parsers
    - sentence parts recognition (noun phrases...)
  - semantics
    - detecting lexical meaning (synonymy/antonymy...)
    - detecting sentence meaning (agent, patient...)
  - machine (aided) translation systems
- commercial products

HLT 3

- resources and tools
  - specific for each language
  - building process starts from the basics for each language
- resources
  - provide fundamental language data (evidence and statistics) for
    - developing other resources
    - building tools

Croatian LT Portal

- public catalogue of
  - institutions
  - projects
  - language
    - resources
    - tools
  - activities

  related to the field of (Croatian) LT
- http://www.hnk.ffzg.hr/jthj

HrLT Portal: target user groups

- target user groups
  - researchers and students dealing with language data
    - humanities
      - linguistics
      - phonetics
      - Croatian language and literature studies
      - translation and interpretation studies
    - IT
      - telematics
      - speech signal processing
      - AI
      - NLP

HrLT Portal: target user groups 2

- IT R&D (companies)
  - natural language interfaces
  - natural language data-indexing
  - full-text search
  - text-compression
  - natural language coded data-mining systems
- telecom/broadcasting companies
  - systems using natural language (e. g. SMS, subtitles)
- translation offices/services
  - MT & MAT

HrLT Portal: structure

- 2 main areas
  - informative component (list at the center)
  - interactive component (forms at the right-hand side)
- navigation bar from left
- e-mail address for contact at the bottom
- structure of the list
  - top: items relevant for Croatian
  - middle: items relevant for other languages
  - bottom: other sites with similar links

Informative component

- institutions and associations
  - Croatian universities/institutes and departments
  - foreign academic institutions and associations
- projects
  - academic projects
  - no non-academic projects
    - company R&D
    - commercial
  - international projects

Informative component 2

- corpora
  - Croatian corpora
    - monolingual
      - Croatian National Corpus (HNK)
        - www.hnk.ffzg.hr
      - 1-million Corpus of Croatian Literary Language
    - bilingual
      - Croatian-English Parallel Corpus
      - Croatian-Slovenian Parallel Corpus
      - Croatian-French Parallel Corpus
  - corpora of foreign languages

Informative component 3

- dictionaries
  - Croatian on-line dictionaries
    - specialized
    - no general dictionary for Croatian
  - on-line dictionaries for other languages
    - general
    - specialized

Informative component 4

- LT tools
  - Croatian tools
    - spelling-checkers (5 items)
    - converters
    - specialized dictionary text editor
  - other tools
    - editors/converters
    - language resources markup (SGML/XML)
    - wordlists, concordancers, text analysis & stats
    - taggers and aligners
    - syntactic parsers and tree-banks
    - semantic nets
    - MT & MAT systems

Informative component 5

- speech processing
  - very scarce data about SP for Croatian
  - MBROLA EU-funded project
    - speech generation, 24 languages (incl. Croatian)
  - Carnegie Mellon University, Pittsburgh
    - Institute for Language Technologies
    - Speech Processing Unit
      - for US Army En-Hr-En speech-to-speech translator tested in Zagreb 2001-04
      - no one LT-aware from Croatia participated
      - in references, use mixture of names: Croatian and Serbo-Croatian

Informative component 6

- conference calls
  - Croatia, Europe, world
- LT glossary
  - 8-lingual with definitions
    - English, French, German, Spanish, Italian, Danish, Finnish, Croatian
- miscellaneous
  - LT projects for people with special needs
  - e-publishing companies
  - links about Croatian language in general

Interactive component

- Direct access and search of HNK
  - 30m = part of HNK with contemporary texts
  - freely searchable
  - result in form of concordance with frequency information about the token
- Discussion area (projected)
  - HrLT forum
- Direct access and search of parallel corpora (projected)
  - Hr-En, Hr-Si, Hr-Fr

Improvements & prospects

- constant updating of data
- new features
  - didactic materials (computational and corpus linguistic courses and tutorials)
- ultimate goals
  - build awareness about necessity of HrLT
  - spread and develop the field of HrLT
  - helping Croatian language to achieve recognition as an independent language
    - by its presence in digital communication channels of the 21st century

Building the Croatian Language Technology Portal

Marko Tadić, Ivana Simeon
(marko.tadic@ffzg.hr, ivana.simeon@ffzg.hr)

Department of linguistics
Institute of linguistics
Faculty of philosophy
University of Zagreb (www.ffzg.hr)

CUC2001, Zagreb, 2000-09-25