The Croatian Language Corpus

Ivo-Pavao Jazbec
Tomislav Stòjanov
@
Institut of Croatian Language and Linguistics

Outline

- What do we want?
- Why do we want it?
- How to get it?
- Let the ball roll!
- Tools
- Target groups
- Looking back to future…

What do we want? (description)

- Well of knowledge about Croatian language
  - Through time (diachronic)
  - In time (synchronic)
  - In space (dialectal)
  - In society (functional styles, professional jargon, terminology)

What do we want? (recommendation)

- Standard/norm – preferred language use:

  - schools, TV/radio (media), publishers

  (Committee for Standardization of Croatian Language)

  - translation

    - Acquis communautaire (EU legal texts)

- Dictionaries & grammars, lexica, language guides

Why?

‘Cause

we

don’t

have it

yet!

How to get it?

- Creation of basic sources:
  - compilation of various corpora of Croatian
- Financing:
  - Primary source:
    - Ministry of Science, Education and Sports
    - …

How to get it? (project)

Institute of Croatian Language and Linguistics
– Dr. Dunja Brozović Rončević
– Dr. Małgorzata E. Ćavar
– Dr. Damir Ćavar
– Ivo-Pavao Jazbec
– Dr. Željko Jozić
– Mr. Tomislav Stojanov
- Dr. Antun Halonja
- Siniša Runjaić
– and many more colleagues in the Institute

How to get it? (project) (2)

- Applied at the Ministry for Science, Education, and Sports of the Republic of Croatia end of 2004
- approved in May 2005 as a research project at the Institute of Croatian Language and Linguistics (initial duration 3 years, yearly evaluation)
- PI: Dr. Dunja Brozović Rončević
- Coordinator of the theoretical and computational linguistics division: Dr. Damir Ćavar

How to get it? (3) (finance)

- A new program applied in spring 2006
  - Various projects, all needing computational (re)sources
  - Among:
    - Computational analysis of CLC, tools and applications (Ćavar) (pending)
    - digitalization of dialectal resources (Jozić) (pending)
    - Digitalization of Old-Croatian written sources (Kapetanović)

How to get it? (4) (legal issues)

- Contrats with:
  - Školska knjiga d.o.o (the biggest publisher in Croatia, broad spectrum of editions/publications – textbooks, children literature, fiction, …
  - HAZU (The Croatian Academy of Sciences and Arts)
  - “Stoljeća hrvatske književnosti”, Matica Hrvatska

Repository

- Creation of an Internet repository for Croatian (Croatian Language On-Line Repository)
  - variants of Corpora
  - Dictionaries (in particular terminological and professonal dictionaries)
  - Online-advisor and information pages for Standard-Croatian
  - Digital maps with texts and recordings from Croatian idioms

Repository (2)

– Digital resources for language technology, e. g. statistical models (N-gram models), wordlists, recordings, grammars

– Basic language technology (various tools, i. e. Part-of-Speech tagger, lemmatizer, parser)

Repository (3)

- Not a single corpus, but a group of corpora
  - corpus of Old-Croatian (till mid of 15th century)
  - corpus of Middle-Croatian (till the beginning of standardization, i. e. mid of the 18th century)
  - Corpus of Kajkavian literary language

Croatian Language Corpus

- Standard-Croatian Corpus (from the beginning of standardization)
- Repressentativness of contemporary Croatian
- Monitor-corpus

Digging the well…

- Available digital formats… (from 1995 partly useful)
- Scanning & OCR
- Tedious correction…
- Person/hours…

Digging the well… (2)

- Current working phase
  - Old-Croatian: scanning and images
  - Middle-Croatian: scanning and images
  - Kajkavian literary language: Scanning and digitization – completed (SoftLex)
  - dialects: digitalisation and transcription of audio material and questionnaires

Digging the well… (3)

- TEI recommendations
- XML based
- Tools that support it
  - philologic

- Copying/xeroxing
  - Scanning and OCR
  - creation of digital images (e. g. TIFF)
  - creation of pdf versions
  - creation of meta-files (title, author, publication details, work description and responsible person)

Digging the well…

- Annotation: Standard is defined by Text Encoding Initiative (TEI) URL: www.tei-c.org
- eXtensible Markup Language (XML) Format standardized types of meta-information and structural elements
- open for specific adaptation/adjustment and expansions

Digging the well…

- XML and TEI XML
  – open and not proprietary standard
  – freely available for processing, convertion, and analyses
  – easy transfer into professional database format

Digging the well…

- step-wise Annotation:
  - Meta-Information
  - Formating information, e. g. italic, bold, gothic
  - Text structure e. g. book parts, chapter, article, paragraphs, sentences, words
  - Linguistic annotation

… never to dry out!

- Linguistic annotation in TEI XML planned:
  - Lemma
  - Part-of-Speech, i.e. morpho-syntactic Annotation but open for extension and specification, e.g.
    - Semantic features
    - specific choice of Part-of-Speech Tags

… never to dry out!

- advantages of the possibility for partial annotation
- step-by-step corpus in-depth building, i.e. inherent text features or linguistic levels
- division of tasks in subtasks and assigning them to experts
- specific quality control after each step

Tools

- Technologies and processes, as well as results:
  - Open standards
  - Freely available resources
  - interchangeable format

Tools (1)

- Philologic (University of Chicago)
  - URL: philologic.uchicago.edu
  - keyword search and context
  - simple statistics
  - collocation analysis
  - possibility of search constraints considering meta-information such as time span, author, location

Tools (2)

- Philologic (advantages)
  - open and free program source
  - access to corpus via Internet with a web-browser of choice
  - offline processing of the corpus and querying speed
  - full TEI XML support and Unicode character encoding

Tools (3)

- Philologic (disadvantages)
  – server-based system, requires highly specialized knowledge for installation, adaptation and adding features
  – offline processing of the corpus and speed of indexing (processor, time and space)
  – limited analysis and tools

Tools (4)

- our approach:
  - Philologic for now
  - adding more possibilities through software development: e.g. functionality, selective limitations, optimization of the working process

How far have we gone

- depth of annotation:
  - text structure to the paragraph level, more ellaborate annotation semiautomatic
  - linguistic components of analysis in development phase:
    - sentence delimitation, morpho-syntactic annotation, syntactic parser with semantic structure, named entity recognizer,…

How far have we gone

- Quantitative:

– more than 71 M words (Tokens)

– more than 650.000 (Types)

Further plans

- analysis of temporal (diachronic) change of quantative lexical (uni- and n-gram models through time) information

  - contextual information of lexical models through vector-space models (lexicological analysis)
- generation of semantic nets (on top of / based on linguistic annotation)

Further plans

- distance and correlation analysis with help of deeper linguistic annotation
- preparation of basic language technologies (lemmatizer, tagger, parser)

Further plans

- data for language tehchnology
  - Training corpora
  - Lexica, n-gram models on the phonological, orthographic, lexical, syntactic level

Online-access: riznica.ihjj.hr

- fulltext-search
- statistical analyses and data (e. g. word-frequency, collocations, keywords in context)

In preparation:

- access to scanned sources
- search over linguistic functions and structures

Target groups

- Project of the dictionary of contemporary Croatian at the Institute
- Approved in 1998
- Ought to be based on the results of the corpus project at another institution approved at the same time
- The other corpus was never satisfactory (neither by size, nor by representativeness)

Target groups (2)

- researchers in linguistics and literature, e. g. from the domains of Croatian and related comparative research
- Committee for the standardization of Croatian appointed by the Minister of Science, Education, and Sports of the Republic of Croatia
- Other national and European institutions

Looking back to future…

- What's the plan for the future
  - Projects (Institute's program, Damir's project)
  - 5-year plan
- Projection in 2011/2012
  - Fully tagged and parsed 180M corpus
  - Manually corrected
  - Parallel corpora

Thank you!