Providing access to language resources through the World Wide Web: the Oslo Corpus of Bosnian Texts

Diana Santos

Text Laboratory, University of Oslo

http://www.tekstlab.uio.no/Bosnian/Corpus.html
Outline of presentation

- the OCBT
- the project
- the language
- the service
  - the general approach
  - the users
  - the queries
- Web serving a corpus: a tripartite view

http://www.tekstlab.uio.no/Bosnian/Corpus.html
The Oslo Corpus of Bosnian Texts

- 1.5 million words
- from 105 different (written) sources
- of 7 different genres
- published in 1989-1995
- available for search in the Web for registered users for non-profit purposes
- parallel interface in Bosnian and in English

http://www.tekstlab.uio.no/Bosnian/Corpus.html
Short history of the OCBT project

First phase
- Project launching by JB Johannessen, 1996
- Text selection by Svein Mønnesland
- Scanning and proofreading the texts by Kemila Basic and Gordana Vranic
- Building a table of contents by Gordana Vranic

Second phase
- Encoding the corpus in the IMS Corpus Workbench
- Creation of a CGI to CQP
- Building a Web service based on the two above
  - monitoring
  - providing help
  - documenting the service
  - adding capabilities

http://www.tekstlab.uio.no/Bosnian/Corpus.html
The Bosnian language

- Bosnian is a South Slavic language
- Before the war (1992-1995), it was called Serbo-Croatian, and two variants (Eastern and Western) were distinguished
- Changes in the language due to the war:
  - reinforcement of the Oriental lexical heritage (Arabic, Persian and Turkish),
  - dominance of the Western variant over the Eastern one in mass-media and everyday speech.

Source: "Recent changes in the Bosnian language as reflected by and documented from the Oslo Corpus of Bosnian Texts", Nedzad Leko, 1998

http://www.tekstlab.uio.no/Bosnian/Corpus.html
The Bosnian language (2)

- 92.5% of Bosnian Muslims declared **Bosnian** as their mother tongue in the **1991** census.
- 67% of Bosnian Serbs stated that Serbian is their language.
- 79% of Bosnian Croats called their language **Croatian**.
- Since Bosnian Muslims represented around 43% of the overall Bosnian population, it means that **before the war more than 40% of citizens regarded Bosnian as their native language**.
- In **1993** a decree was issued by the Bosnian government in which the Iekavian dialect was recognised as the official language of the state; **one of three names for the language may be used: Bosnian, Croatian, or Serbian**; both scripts (Latin and Cyrillic) are proclaimed equal. (…) in school certificates the mother tongue should be labelled as Bosnian, Croatian, or Serbian according to each student’s wish.  
  
  ![Bosnia and Herzegovina Map](http://www.tekstlab.uio.no/Bosnian/Corpus.html)
Bosnian native speakers

- 2.7 millions of Bosnians living in Bosnia now
- 1 million Bosnians around the world due to the war
- A large emigrant community
- Bosnian+Croatian+Serbian native speakers in the area of former Yugoslavia: 19 millions

http://www.tekstlab.uio.no/Bosnian/Corpus.html
Giving world access to corpora

- using the Web as infrastructure
- minimizing requirements from the user’s side
- providing remote service / maintenance / feedback instead of “giving away” the corpus

http://www.tekstlab.uio.no/Bosnian/Corpus.html
The OCBT users’ map

- 39 registered users
- from 15 different countries
  - Germany
  - Switzerland
  - USA
  - England
  - Spain
  - Australia
  - Norway
  - Bosnia

http://www.tekstlab.uio.no/Bosnian/Corpus.html
OCBT statistics

- From 24/12 to 5/5/98
  - 3913 queries
- Since counting results
  - 1962 queries
  - 1178 cases where some result of query existed
  - 43 cases asking for distribution only
  - 26 queries with random selection

http://www.tekstlab.uio.no/Bosnian/Corpus.html
OCBT technical description

- a Web server
  - no special requirements for the user’s browser
    - plain HTML
    - hopefully with support for ISO-Latin2
- a Web interface programmed in Perl 5.0
- the IMS Corpus Workbench in a Solaris workstation

http://www.tekstlab.uio.no/Bosnian/Corpus.html
Monitoring examples

Looking for the distribution of voz.* (voz: train N, voziti: drive V)

On 8 Jan 1998, User1 searched for "voz.*" with 366 results
On 8 Jan 1998, User1 searched for voz with 70 results
On 8 Jan 1998, User1 searched for voza with 19 results
On 8 Jan 1998, User1 searched for vozu with 10 results
On 8 Jan 1998, User1 searched for vozom with 8 results
On 8 Jan 1998, User1 searched for vozovi with 5 results
On 8 Jan 1998, User1 searched for vozova with 8 results
On 8 Jan 1998, User1 searched for vozovima
On 8 Jan 1998, User1 searched for vozove with 2 results
On 8 Jan 1998, User1 searched for vozovima
On 8 Jan 1998, User1 searched for vozovima
On 8 Jan 1998, User1 searched for Vozi with 9 results
On 8 Jan 1998, User1 searched for VOZ
On 8 Jan 1998, User1 searched for Voza
On 8 Jan 1998, User1 searched for VOZA
On 8 Jan 1998, User1 searched for Vozu
On 8 Jan 1998, User1 searched for VOZU
On 8 Jan 1998, User1 searched for Vozom with 1 results
On 8 Jan 1998, User1 searched for VOZOM
On 8 Jan 1998, User1 searched for Vozovi with 1 results
On 8 Jan 1998, User1 searched for VOZOVI
On 8 Jan 1998, User1 searched for VOZOVA
On 8 Jan 1998, User1 searched for Vozova with 1 results
On 8 Jan 1998, User1 searched for Vozovima
On 8 Jan 1998, User1 searched for VOZOVIMA
On 8 Jan 1998, User1 searched for Vozove
On 8 Jan 1998, User1 searched for VOZOVE

http://www.tekstlab.uio.no/Bosnian/Corpus.html
Monitoring examples

On 17 Feb 1998, User3 searched for Sarajevo with 636 results

On 23 Feb 1998, User4 searched for "Izet.*" "S.*" with 7 results

On 24 Feb 1998, User4 searched for "bosan.*" "jezik.*" with 97 results

On 13 Mar 1998, User1 searched for "boj.*".*-(ale) with 138 results (KWIC)
On 13 Mar 1998, User1 searched for "se""boj.*".*-(ale) with 15 results (KWIC)

On 18 Mar 1998, User2 searched for "zna.*" [[0,3] "bi.*" (DISTFORM)
On 18 Mar 1998, User2 searched for "zna.*" [[0,3] "bi.*" with 318 results (KWIC)
On 18 Mar 1998, User2 searched for "zna.*" [[0,3] "bi.*" (DISTSOURCE)

On 23 Mar 1998, User5 searched for hajvan.* with 11 results (KWIC)
On 23 Mar 1998, User5 searched for Hajvan.* with 1 results (KWIC)
On 23 Mar 1998, User5 searched for hjavan.* with 11 results (KWIC)

On 3 Apr 1998, User4 searched for {word="kak.*" & ori="E/LA.*"} with 128 results (KWIC)

On 10 Apr 1998, User4 searched for stotine with 143 results (KWICDISTFORM)
On 10 Apr 1998, User4 searched for "*i" stotine (KWICDISTFORM)
On 10 Apr 1998, User4 searched for "*i" stotine with 44 results (KWICDISTFORM)
On 10 Apr 1998, User4 searched for "*i" stotine with 46 results (KWICDISTFORM)
On 10 Apr 1998, User4 searched for "*i" stotine with 35 results (KWICDISTFORM)
On 10 Apr 1998, User4 searched for "*i" stotine with 10 results (KWICDISTFORM)

On 22 Apr 1998, User2 searched for "zna.*""*j""""i"" with 6 results (KWIC)
On 22 Apr 1998, User2 searched for "znali""smo""""i" (KWIC)
On 22 Apr 1998, User2 searched for "znali""su""""i" with 2 results (KWIC)
On 22 Apr 1998, User2 searched for "znali""su""""i" with 2 results (KWIC)
On 22 Apr 1998, User2 searched for "znalo""se""""i" with 1 results (KWIC)
On 22 Apr 1998, User2 searched for "znalo""se""""i" with 3 results (KWIC)

On 22 Apr 1998, User2 searched for "zna.*""se""""i" with 891 results (KWIC)
On 22 Apr 1998, User2 searched for "znalo" with 35 results (KWIC)
On 22 Apr 1998, User2 searched for "znali" with 213 results (KWIC)
On 22 Apr 1998, User2 searched for "znam" (KWIC)
On 22 Apr 1998, User2 searched for "znamo" with 111 results (KWIC)
On 22 Apr 1998, User2 searched for "znato" with 112 results (KWIC)
On 22 Apr 1998, User2 searched for "znala" with 227 results (KWIC)
On 22 Apr 1998, User2 searched for "znam" (KWIC)
On 22 Apr 1998, User2 searched for "znas" (KWIC)
On 22 Apr 1998, User2 searched for "zna" with 312 results (KWIC)
On 22 Apr 1998, User2 searched for "zna" with 6 results (KWIC)

http://www.tekstlab.uio.no/Bosnian/Corpus.html
A tripartite view of corpus serving

- An interface is more than a window to something that already exists
- It can add functionalities, serve as a tool, be useful as an educational object

http://www.tekstlab.uio.no/Bosnian/Corpus.html
Evaluation of Web corpus service

- ease of use
- documentation and help
- query power
- speed
- display of results
- technical appropriateness

http://www.tekstlab.uio.no/Bosnian/Corpus.html
What was learned

- It pays to automate corpus creation
- User support and maintenance take the lion’s share of developer’s time
- Monitoring user’s attempts in order to provide improvements is a way to stay in tune with the users community
- Not every linguist knows what to do with a corpus: need to raise corpus awareness

http://www.tekstlab.uio.no/Bosnian/Corpus.html
### Results of searching the Oslo Corpus of Bosnian Texts

**Request string:** "\[v\|o\|e\|e\]".

**Request for:** Distribution of forms

<table>
<thead>
<tr>
<th>Form</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>voz</td>
<td>70</td>
</tr>
<tr>
<td>vozl</td>
<td>47</td>
</tr>
<tr>
<td>vozlž</td>
<td>30</td>
</tr>
<tr>
<td>vozlžč</td>
<td>22</td>
</tr>
<tr>
<td>vozlžo</td>
<td>20</td>
</tr>
<tr>
<td>vozlžača</td>
<td>19</td>
</tr>
<tr>
<td>vozlžačo</td>
<td>17</td>
</tr>
<tr>
<td>vozlžačož</td>
<td>13</td>
</tr>
<tr>
<td>vozlžačožč</td>
<td>13</td>
</tr>
<tr>
<td>vozlžačožčo</td>
<td>10</td>
</tr>
<tr>
<td>vozlžačožčož</td>
<td>9</td>
</tr>
<tr>
<td>vozlžačožčožč</td>
<td>9</td>
</tr>
<tr>
<td>vozlžačožčožčo</td>
<td>8</td>
</tr>
<tr>
<td>vozlžačožčožčož</td>
<td>8</td>
</tr>
<tr>
<td>vozlžačožčožčožč</td>
<td>7</td>
</tr>
<tr>
<td>vozlžačožčožčožčo</td>
<td>6</td>
</tr>
<tr>
<td>vozlžačožčožčožčož</td>
<td>6</td>
</tr>
<tr>
<td>vozlžačožčožčožčožč</td>
<td>5</td>
</tr>
<tr>
<td>vozlžačožčožčožčožčo</td>
<td>4</td>
</tr>
<tr>
<td>vozlžačožčožčožčožčož</td>
<td>4</td>
</tr>
<tr>
<td>vozlžačožčožčožčožčožč</td>
<td>3</td>
</tr>
<tr>
<td>vozlžačožčožčožčožčožčo</td>
<td>3</td>
</tr>
<tr>
<td>vozlžačožčožčožčožčožčož</td>
<td>2</td>
</tr>
</tbody>
</table>

387 instances found.