

# **What do users want from an on-line dictionary: A seven-year usage study of an e-dictionary**

Vlado Keselj, Dalhousie University  
Tanja Keselj, Korlex Software



# Presentation Overview

- electronic dictionaries: design issues
- rechnik.com: an English-Serbian on-line bi-directional dictionary
- directions for further improvements
- usage statistics and trends
- query content analysis
- analysis of un-answered queries
- conclusions

# Electronic Dictionaries

- Two aspects of e-dictionary design:
  1. ED as marked-up text
    - encode classical dictionary using XML-based markup and provide elaborate searching capabilities [project OED]
  2. Knowledge-base structure of ED
    - provide manageable and efficient structure of e-dictionary
    - time-stamping, variation mark-up

# Creating a on-line Web dictionary

- typical approaches:
  - RDMS support, e.g. MySQL
  - CGI using LAMP framework  
(Linux+Apache+MySQL+Perl/php/Python), Ruby on rails
- using a database server seems to be an overkill
- YourDictionary.com lists on-line dictionaries for 294 languages (R.Beard, Bucknell U.), including two for sign languages

# A Simple Approach

- A simple approach in rechnik.com:

Enter a word or phrase

Search

Examples: student, word, djak, re"c,  
comput# for all words starting with comput

Results:

abash, bewilder, confound, confuse = zbuniti, zaplesti, zavesti, posramiti, pobrkati

- using: simple line-based indexing with Perl

# Recnik.com project description

- since 1999, English-Serbian bilingual dictionary, seems to be one of the oldest still active
- Some other dictionaries can be found at YourDictionary.com and Google.com
- e.g.:  
krstarica.com,  
[www.public.asu.edu/~dsipka/rjeynici.html](http://www.public.asu.edu/~dsipka/rjeynici.html)

# Design features

- text-based format, one meaning = one line
- format translatable to XML (TEI)
- Web on-line search interface is available
- interface to LaTeX-based printed version
- format details
  - transliterated phonetic description
  - encoded ekavian/ijekavian dialect variations, e.g:  
ml{ij}eko, prim{j}eri, d{e|i}o, ht{e|i}o, etc.

# Dictionary Structure

- inspired by the WordNet structure:
  - one meaning = one entry
- example
  - abash [ˈæbʰæːs], bewilder [biwˈildˌer], confound \ [kʰanfˈaund], confuse [kˌenfjˈuːz] = \
  - :v zbuniti, zapelesti, zavesti, posramiti, ::coll pobrkati,
  - :eg too much choice can bewilder a small \
  - child = prevelik izbor moːze zbuniti malo d{ij}ete

# Comparative TEI Structure

- translatable into the TEI structure, e.g:  

```
<entry key="bewilder"><form>  
<orth type='hw'>bewilder</orth><pron>bl"wild@(r)  
</pron></form>  
<gramgrp><pos>vrt</pos></gramgrp>  
<sense orig='sem'><trans><tr>zbuniti</tr>,  
  <tr>zaplesti</tr>, <tr>pobrkati</tr></trans>  
<eg><quote>too much choice can bewilder a small  
  child</quote><trans><tr>prevelik izbor mo"ze  
  zbuniti malo d{ij}ete</tr></trans>  
</eg></entry>
```

# Current state

- about 75 000 lexemes
- 20-30 000 synsets (e.g., lines, meanings)
- 7-year usage experience
- future improvements:
  - add more...
  - better support for TEI format
  - include a lemmatizer
  - detection and correction of typos
  - translation of complex phrases

# Questions

- How do we know what improvement would be useful?
- Are finite-state methods a desirable tool?

# Study Objectives

- to explore query data in the 7-year period:
- frequency, query statistics
- query complexity, trend
- query content, categorization
- structure and categorization of un-answered queries
- is the use of finite-state methods a promising and desired additional methodology?

# Usage Statistics

Year	Avg.visits per day	Avg. time b/w visits	Len. of the longest query
1999	106	13m 34s	953
2000	249	5m 47s	710
2001	402	3m 34s	1556
2002	662	2m 10s	2492
2003	1018	1m 25s	4958
2004	2158	40s	1249
2005-	2-3K	30-40s	-

# The most common queries

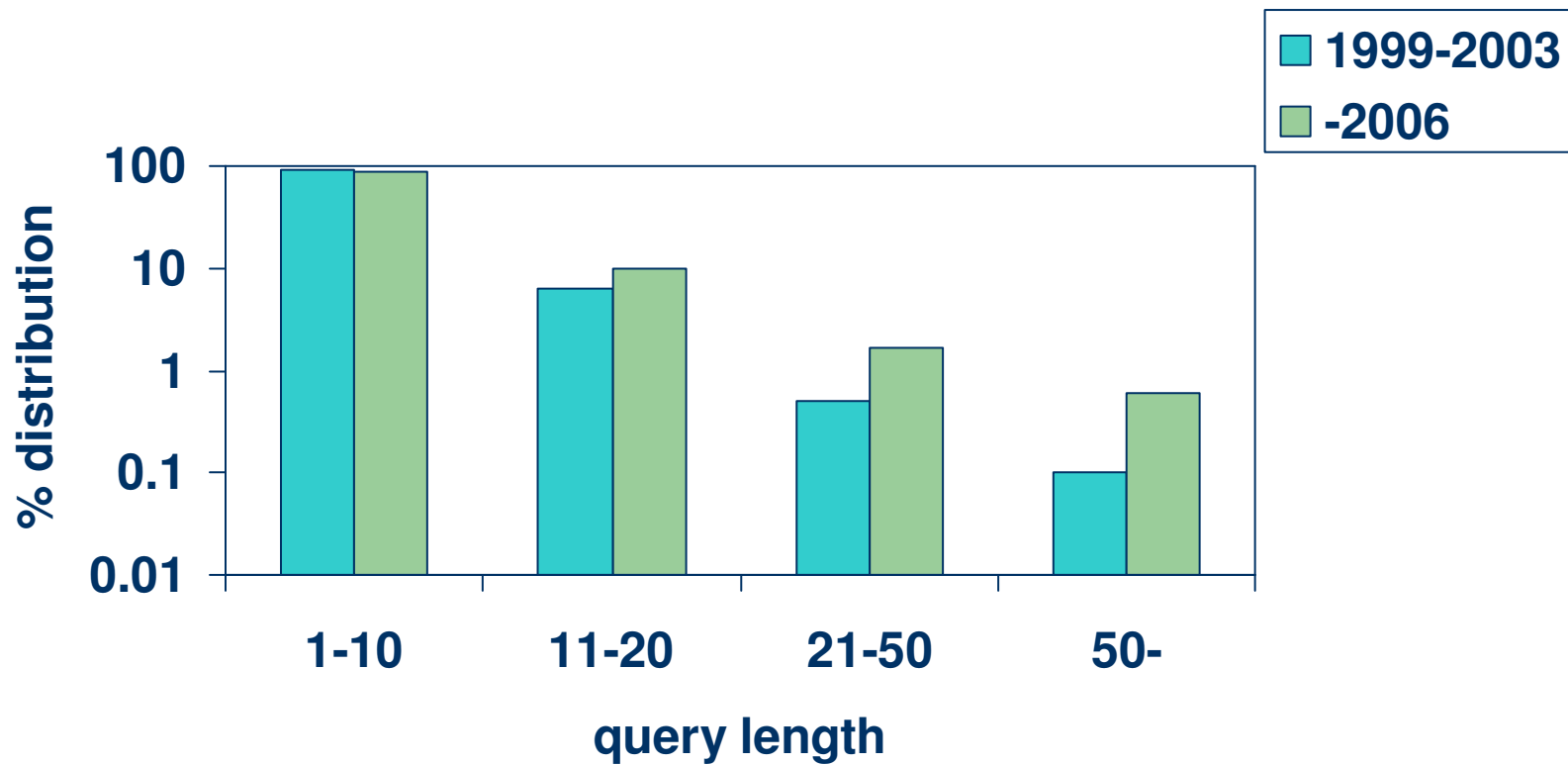
1999	2000	2001	2002
0.6 hello	0.7 love	0.6 love	0.5 hello
0.6 love	0.6 hello	0.5 hello	0.5 love
0.4 you	0.4 you	0.4 you	0.4 you
0.4 devojka	0.3 good	0.2 good	0.2 i
0.2 i	0.2 f* (en)	0.2 i	0.2 good
0.2 k* (sc)	0.2 i	0.2 devojka	0.1 f* (en)
0.2 djevojka	0.2 I	0.2 f* (en)	0.1 thank you
0.2 djak	0.2 devojka	0.2 thank you	0.1 happy
0.2 f* (en)	0.2 are	0.2 happy	0.1 beautiful
0.2 word	0.2 thank you	0.2 I	0.1 I

# The most common queries

2003	2004	2005	2006
0.5 hello	0.4 HELLO	0.5 hello	0.4 hello
0.5 love	0.3 love	0.5 love	0.4 love
0.2 you	0.2 YOU	0.3 you	0.3 I love you
0.2 good	0.2 I	0.3 i	0.3 student
0.2 i	0.2 I love you	0.2 student	0.2 you
0.2 thank you	0.1 good	0.2 i love you	0.2 how
0.2 f* (en)	0.1 HOW ARE YOU?	0.2 good	0.2 how are you?
0.2 beautiful	0.1 student	0.2 how are you?	0.2 i
0.1 are	0.1 thank you	0.2 happy	0.2 djak
0.1 i love you	0.1 hello	0.2 djak	0.2 good



# Query Lengths in Characters



# Most frequent query types

- methodology:
  - collected 50 most frequent queries over each quarter (3-month period)
  - summed over each year
  - divided into “meaningful” syntactic-semantic categories
  - first into 35 finer-grained categories
  - clustered into larger groups

# Most frequent queries: 2006

18.70 %	Love communication related
18.30 %	Greetings and holidays
15.76 %	Pronouns (personal, interrogative)
9.04 %	Education domain
8.70 %	Frequent functional words (aux, clit.)
6.60 %	Frequent adverbs
3.76 %	Frequent verbs
3.64 %	Frequent nouns
2.62 %	Sex/obscenities related

# Most frequent queries: 2005

23.37	Greetings and holidays
20.28	Pronouns
11.70	Frequent functional words
9.31	Love related
8.23	Frequent nouns
5.00	Frequent adverbs
4.03	Education domain
3.84	Sex/obscenities
3.29	Frequent verbs
	3.18 incorrect encoding

# Most frequent queries: 2004

21.11 %	Greetings and holidays
15.71 %	Pronouns
15.21 %	Frequent functional words
10.81 %	Frequent nouns
9.66 %	Love related
5.29 %	Frequent adverbs
4.71 %	Frequent verbs
4.35 %	Sex/obscenities
0.54 %	Education domain
	Unusually frequent: pelinkovac

# Most frequent queries: 2003

15.36 %	Frequent adverbs
15.12 %	Pronouns
13.07 %	Greetings and holidays
12.69 %	Frequent functional words
9.83 %	Frequent nouns
8.54 %	Love related
6.59 %	Frequent verbs
5.14 %	Sex related
2.93 %	Frequent adjectives
0.78 %	Education domain

# Query category examples

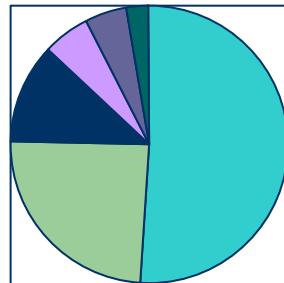
- A large group is greetings, which includes:
  - good day, hello, hi, ...
  - please, thank you
  - holidays and holiday greetings
- frequent nouns include significant portions:
  - household, car, family
- frequent functional words include (2006):
  - 25% da/ne, 25% do/be/have, 50% other
- pronouns: personal + interrogative

# Unanswered Queries

- some are legitimate entries, missing from dictionary – periodically added
- the rest can be divided into:
  - incorrect queries (typos, errors, other languages)
  - encoding problems
  - too long queries, expected translations of whole phrases and sentences
  - morphological analysis required (not a lemma)

# 2006 Queries with no answers

50.72%	Morphological variations
24.55%	Phrases, valid candidates for inclusion
11.55%	Queries with typographical errors
5.60%	Long passages, not feasible candidates
5.05%	Unsupported encoding
2.53%	Proper nouns



- Morphological variations
- Phrases
- Typos
- Passages
- Unsupported encoding
- Proper nouns

# Some examples

- Encoding examples
  - valid transliteration: djak, gre”ska, ku’ca, vi”se
  - UTF-8, &#1057; encoded, Latin-2, ISO
- Proper nouns:
  - george, tivat, Milosevic, bijeljina
- Incomplete coverage:
  - klizi”ste, behar, “zalfija, heklanje, mravojed

# Valid phrases

- examples: of course, see you, volim te
- what is your name, you are welcome
- have a nice day
- coat of arms, Postovani Prijatelji
- tip pretplatnika
- date of birth, I am, I like you, see you
- what is your name

# Incorrectly entered queries

- dont, Whats up?, ori
- Z"eleti
- de, desese, metaphor, jebiga, neznam
- thru, Yadrans, Jel
- volimte, matimu, goodnight
- fala, opshtina, slatak, cower, ljepa, hrpski
- vi"se, jel, uza"sno, ljepa

# Morphological variations

- bolje, imam, javi, mila, nemoj
- radis, smo, zelim, zidinama
- bila, nisam, reci, bolje, slusam
- kisses, eyes, keywords, refugees, fibroids
- occurred
- o”ciju, “cujemo, drugarice
- frequent 1<sup>st</sup> person singular or imperative form

## Long phrases, examples

- Sometimes whole messages (page or two)
- ja sam dobro. Kako si ti?
- trgovina odje'ce i obu'ce
- Boli me glava. Vrtili mi se u glavi.
- Kosu nisam odsekla.

# Conclusions (1)

- usage of an on-line dictionary is more similar to a typical search engine usage than a classical dictionary
- on-line dictionary should make a good support for common communicative words and phrases
- users probe and expect that more complex queries are translated

## Conclusions (2)

- if a goal is to give a good support to a maximal number of queries then:
  - support for common categories is necessary (greetings, pronouns and functional words, human relationship domain)
  - support for basic morphological variations:
    - e.g. at least first person singular forms and imperative
    - noun plural for English
    - frequent functional words, clitics, and similar

# Future work

Further analysis:

- automatic classification and clustering of queries

Dictionary improvements:

- trained lemmatizer
- trained typographical correction
- encoding detection
- narrow grammar for complex queries