Online Style Guide for Slovene

Session 4

History

- (written) language codification or standardization of Slovene
- "good writing" – Slo. pravopis, Ger. Rechtschreibung
- the first one published in 1899
- the most recent one published in 2001
- rules & dictionary

Ortographic dictionaries

- emphasis on orthographically challenging vocabulary
- last orthography guide
  - published in 2001 in printed form
  - on CD-ROM in 2003
  - available online since 2010 (Neva, Termania, ASP+)
- digital version replicates the printed one
- rules available as a PDF document

World of Print

- printing process (in Slovenia):
  - the author
  - the editor (= publishing house)
  - the proof-reader
  - a language specialist called "lektor"
- "lektors" are responsible for the compatibility of published texts with the language norm or standard

The Times They Are A-Changin'

- print cycle is now more or less broken
  - the possibility to publish texts online without the assumed or axiomatic interference of third parties
  - the time needed from the creation of the text to its publication has been reduced to just a few seconds
  - numerous genres previously reserved for private communication are now part of the public sphere

Codification in digital environment

- codification-related language help currently comes from two basic sources
- spelling or grammar checkers
  - replace the proof-reader
- online portals, dedicated forums, social networks, search engines
  - providing consultation or feedback from
    - peer communities
    - official bodies responsible for language codification

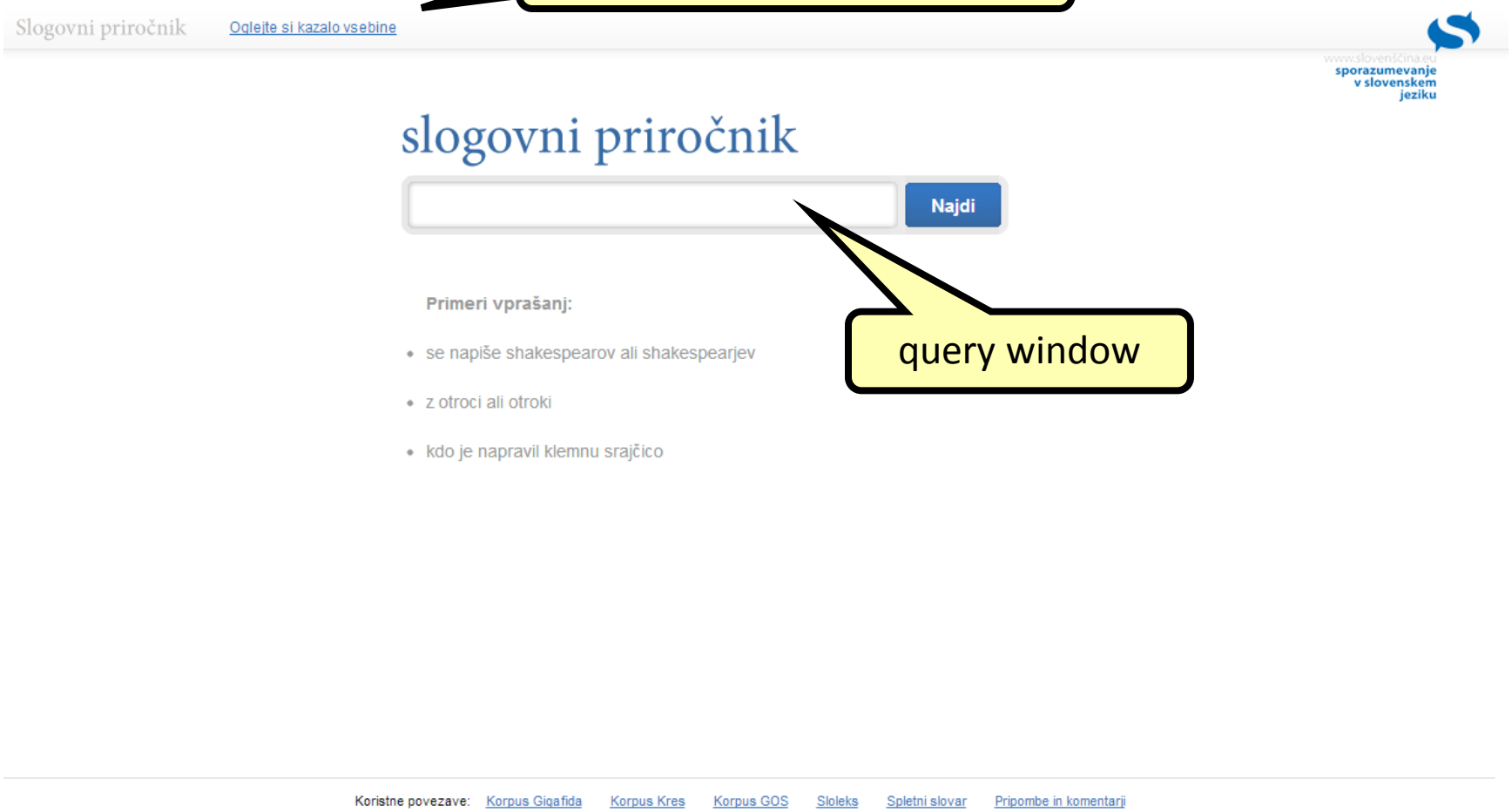
Ortography Guide web portal

- answer to the question of how language codification should be presented in the digital (web) environment of the 21st century
- providing standardized explanations of the most frequent problems with language or (more narrowly) spelling and orthography

- DEMO: http://slogovni.slovenscina.eu/

Prva stran

index of language problems



List of language problems

- 700 detected language problems functioning as the central database
- analysis of traditional orthography guides, text corpora and web forums specialized in language (web crawling)
- special data mining procedures
  - lists of variant forms of words where speakers (or writers) of Slovene falter

Ontology

LABEL	CATEGORY
D	word-formation
D1	adjectives
D1a	possessive adjectives from names of masculine gender
D1a1	from names ending in vowels
D1a1a	from names ending in -a
D1a1b	from names ending in unpronounced -e
D1a1c	from names ending in -y

Luka – Lukov / Lukin
Matija – Matijev / Matijin
Strniša – Strnišev / Strnišin

Wilde – Wildov / Wildeov
Sartre – Sartrov / Sartreov

Sarkozy – Sarkozyjev / Sarkozyev
May – Mayev / Mayjev

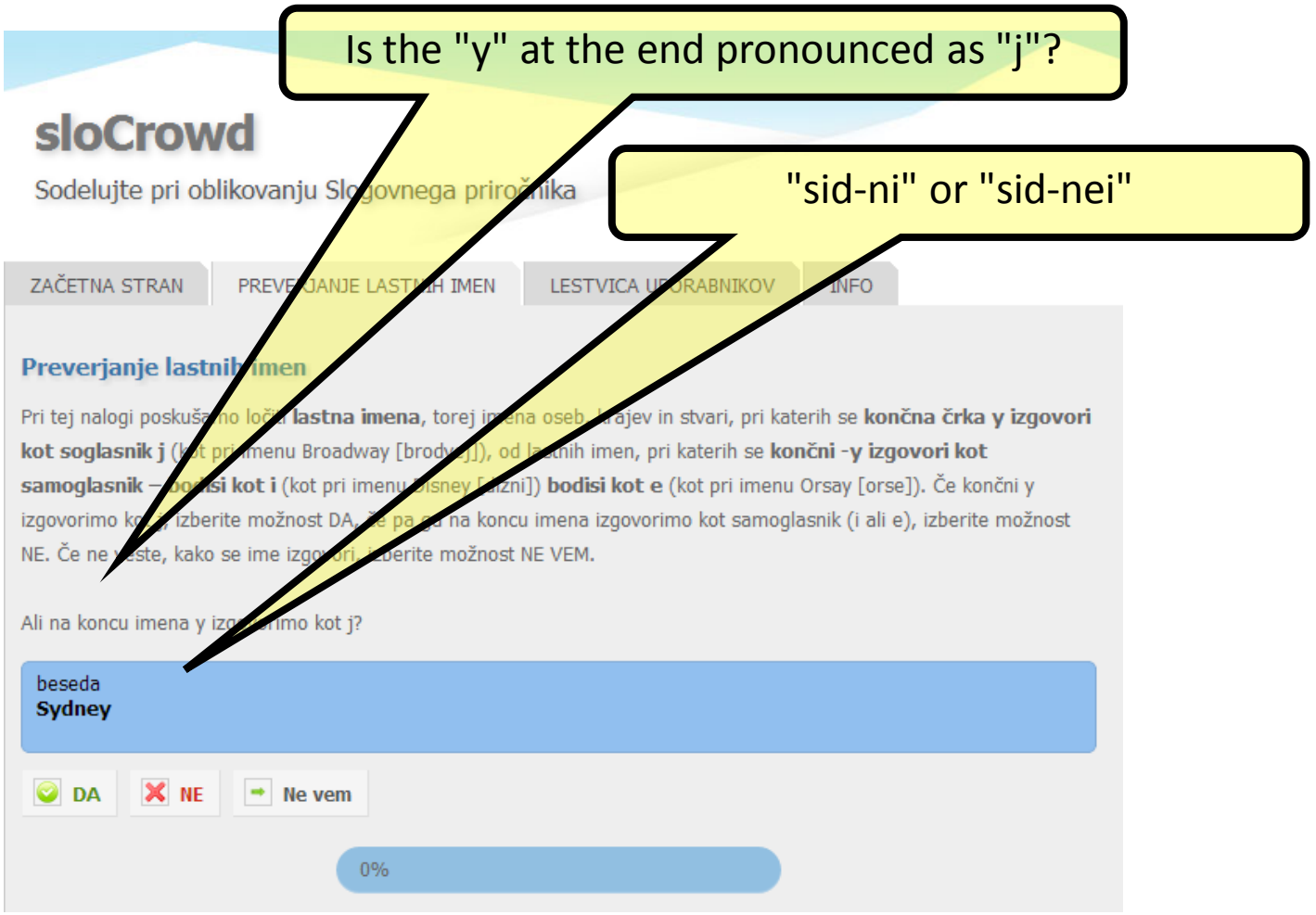
Data mining

Declension of Slovene names of masculine gender with the ‘unsteady vowel’ (C1a3a)

koren	en	n	LOG(D2)	LOG(E2)	score_1	score_2
Klem	1843	3839	3,75	4,48	**20,01**	**1,26**
Niels	164	120	2,45	3,40	**17,57**	**0,73**
Berg	208	375	2,77	3,22	**14,63**	**0,71**
Natlač	223	147	2,57	3,07	**17,26**	**0,67**
Robb	163	333	2,70	3,21	**14,34**	**0,65**
Gold	37	29	1,82	3,80	**17,79**	**0,61**
Gall	105	148	2,40	2,89	**13,42**	**0,60**
Franz	117	114	2,36	2,72	**14,19**	**0,58**
Bjorndal	112	163	2,44	2,75	**18,29**	**0,58**
Patt	85	113	2,30	2,77	**12,94**	**0,56**
Jens	138	60	2,30	3,15	**13,86**	**0,51**

Declension of foreign names of masculine gender with the ‘unsteady vowel’ (C1a3b)

Crowd-sourcing



Three-layered configuration of answers

- The **short answer** consists of text in XML format which can generate a formulaic textual answer with relevant statistical data from the corpus and the lexicon
- Each identified problem in the ontology receives **one long answer** which is written in HTML format and included in the central database.
- "**For enthusiasts**" provides links to scholarly works related to the particular problem

Short answer

- Using data directly from
- Gigafida corpus
  - web concordancer (http://www.gigafida.net/)
  - 1.2 billion word corpus (tagged & lemmatized)
- Sloleks lexicon
  - web service (http://eng.slovenscina.eu/sloleks)
  - 100,000 lemmas, 2.8 million word forms
  - normative information

Short answer – XML structure

- <!-- variant 1: FOUR, standard-12, non-standard-34 -->
- <text var="S00.S00.N00.N00" graph="1234">
- The graph shows the data about the use of word forms <word id="1"/>, <word id="2"/>, <word id="3"/> and <word id="4"/> in the Gigafida corpus. Word forms in blue color are standard, those in grey are not compatible with the current standard of written Slovene.</text>
- </tekst>

Element <beseda> will be replaced by the keyword recognized in the query

this answer will be shown if there are four possible word forms of the recognized keyword in Sloleks

four word forms will be shown on the graph, two standard, two non-standard

```
<LexicalEntry  id="LE_S_foyer">
        <feat att="besedna_vrsta" val="samostalnik"/>
        <feat att="vrsta" val="občni"/>
        <feat att="spol" val="moški"/>
        <Lemma>
           <feat att="zapis_oblike" val="foyer"/>
        </Lemma>
        <WordForm>
           <feat att="število" val="ednina"/>
           <feat att="sklon" val="rodilnik"/>
           <FormRepresentation>
                <feat att="zapis_oblike" val="foyerja"/>
                <feat att="norma" val="variantno"/>
                <feat att="pogostnost" val="15"/>
           </FormRepresentation>
           <FormRepresentation>
                <feat att="zapis_oblike" val="foyera"/>
                <feat att="norma" val="variantno"/>
                <feat att="pogostnost" val="1"/>
           </FormRepresentation>
        </WordForm>
         <RelatedForm>
           <feat att="idref" val="LE_P_foyerov"/>
        </RelatedForm>
        </WordForm>
</LexicalEntry>
```

unique ID

lemma

word form

normative status

frequency

related form
(word formation)

Sloleks and the ontology

<LexicalEntry id="**LE_S_Klemen**" xmlns:d="urn:LEKSIKON_SSJ">
<feat att="besedna_vrsta" val="**samostalnik**" />
<feat att="**SP2001**" val="samostalnik" />
<feat att="vrsta" val="**lastno_ime**" />
<feat att="spol" val="**moški**" />
<feat att="SPSP" val="**C1a3a**" />
<Lemma>
<feat att="zapis_oblike" val="**Klemen**" />
</Lemma>
<...>

Normative category from the ontology

Long answer

- fixed (one answer per category)
- standardized (length, structure, etc.)
- user friendy (language understood by laymen)
- linked (corpora, dictionaries, Wikipedia, etc.)
- with additions
  - lists of words belonging to the same category
  - terminology in pop-up windows
- in HTML format

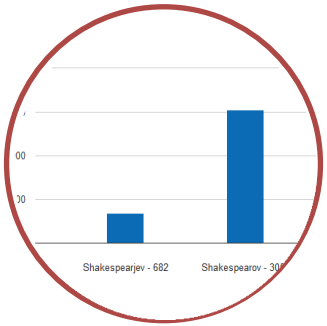
"For enthusiasts"

- official reference works in digital form
- other scanned & OCRed reference books
- other resources (scientific articles, etc.)

- if possible, opened on the relevant page

So how does the portal work?

Parse the user query → Compare the result with Sloleks → Prepare data for the short answer → Visualize the answer on the portal

Strnišev ali Strnišin? → D1a1a → Strnišev / Strnišin → 

Short demo?

- http://slogovni.slovenscina.eu/