The IMP project: developing resources for historical Slovene

Tomaž Erjavec

Dept. of Knowledge Technologies
Jožef Stefan Institute

Ljubljana

First ENeL workshop
September 29 2014
Bled

Overview

1. Project background
2. Digital library
3. Hand-annotated corpus
4. Lexicon
5. Annotation tool
6. Conclusions

Background

Improving Access to Text
IMPACT

Google

1. **AHLib** project (2004–08)
   - *Deutsch-slowenische/kroatische Übersetzung 1848 – 1918,*
     AAS / KFU (prof. Erich Prunč) + JSI
2. **EU IP IMPACT** (2010–2011)
   - *Improving Access to Text*,
     NUK (Alenka Kavčič Čolić, Ines Vodopivec) + JSI
3. **Google award** (2011 + 2012)
   - *Computational models for historical Slovene*,
     ZRC SAZU (Matija Ogrin) + JSI
   - Funding for Wikisource (Miran Hladnik)

The Texts and Goals of the Projects

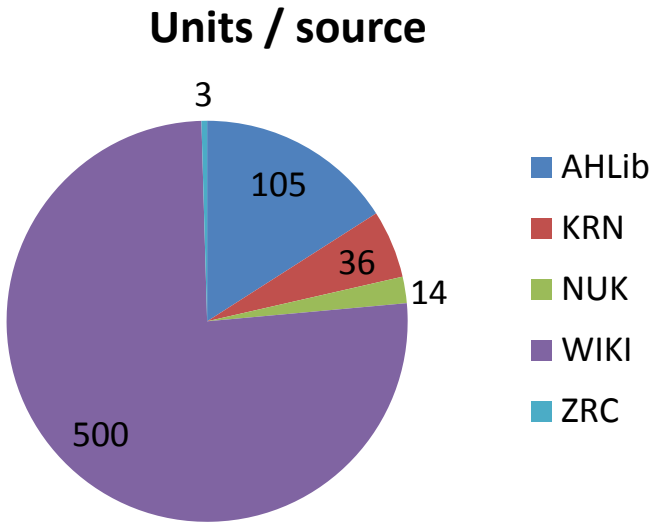
- AHLib
  - Slovene books 1848 – 1918 translated from German
  - Develop a corpus to study translation processes
- IMPACT
  - Slovene books + 1 newspaper from 18th and 19th centuries
  - Proof-read texts and facsimiles for improving OCR and lexicon of historical words for improving IR
- Google:
  - Two samples of very old Slovene books + Wikisource Slovene literary classics from (mainly) 1850 - 1918
  - develop computational models for historical Slovene for better language technologies

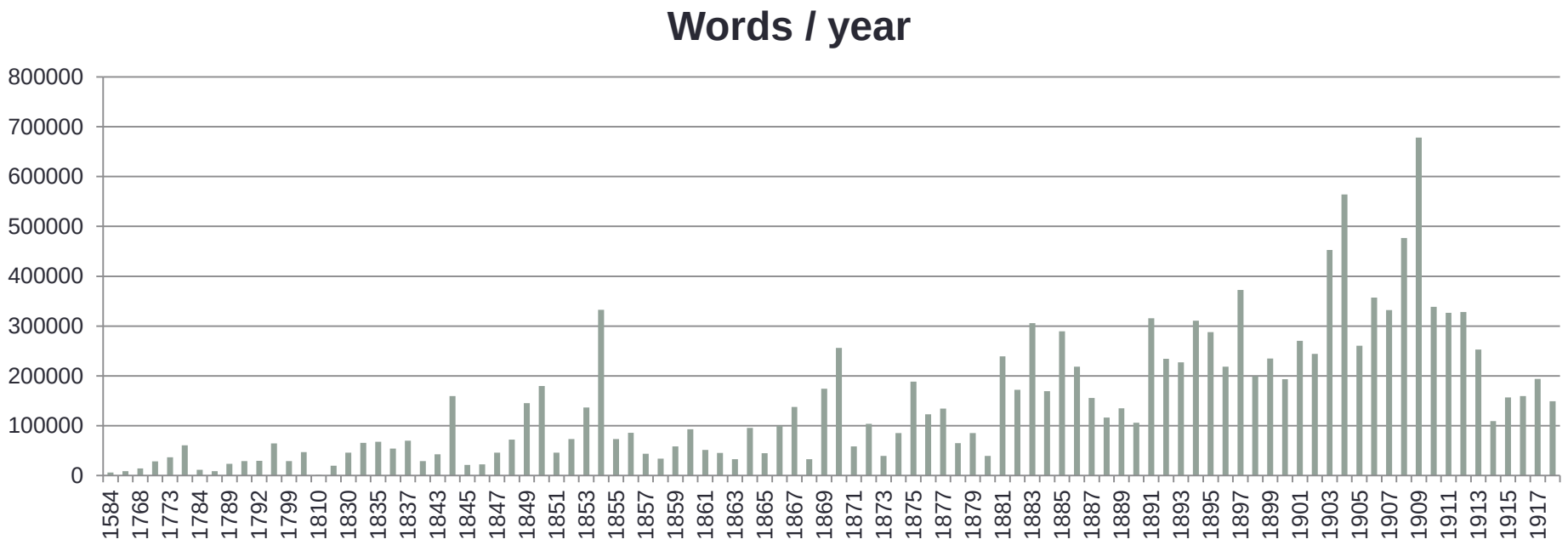
IMP Digital Library

- Each text consists of:
  - **meta-data**
  - **facsimile**
  - **proof-read transcription**
- All texts are uniformly encoded according to *TEI Guidelines for Electronic Text Encoding and Interchange*, **TEI P5**
- No problems with further use or dissemination (**CC-BY**)

IMP DL by source

Source	Units	Pages	Words
AHLib	105	12.854	2.783.569
KRN	36	555	551.599
NUK	14	2.428	338.874
WIKI	500	30.296	10.761.913
ZRC	3	75	20.631
∑	**658**	**46.208**	**14.456.586**





Distribution of DL texts

	Texts		Pages		Words	
	658	100.0%	44,951	100.0%	14,473,005	100.0%
1584	1	0.2%	8	0.0%	5,880	0.0%
1695	1	0.2%	27	0.1%	8,619	0.1%
1768–1799	14	2.1%	2,345	5.2%	379,011	2.6%
1800–1849	27	4.1%	3,653	8.1%	887,589	6.1%
1850–1899	368	55.9%	22,647	50.4%	7,400,493	51.1%
1900–1918	247	37.5%	16,271	36.2%	5,753,403	39.8%
book	275	41.8%	29,157	64.9%	7,084,733	49.0%
magazine	309	47.0%	13,850	30.8%	6,085,893	42.0%
newspaper	63	9.6%	1,674	3.7%	1,175,004	8.1%
manuscript	11	1.7%	270	0.6%	89,365	0.6%
religious	32	4.9%	4,889	10.9%	988,694	6.8%
nonfiction	127	19.3%	8,661	19.3%	3,004,279	20.8%
play	21	3.2%	2,379	5.3%	360,153	2.5%
poetry	3	0.5%	906	2.0%	191,138	1.3%
prose	475	72.2%	28,116	62.5%	9,890,731	68.3%
original	547	83.1%	32,316	71.9%	11,583,141	80.0%
translation	111	16.9%	12,635	28.1%	2,851,854	19.7%

Annotation on each unit

- Meta-data (teiHeader):
  - id, responsibility, extent, availability …
  - basic bibliographic information (two titles: original, modern)
  - taxonomy: medium (manuscript, book, magazine, newspaper), text type (fiction, non-fiction, religious), translation status (original, translated)
  - tag usage, revision description
- Facsimile:
  - images in several sizes, each page break linked to facsimile
- Text structure:
  - divisions, headings, lists, tables, notes, poems, figures, line breaks…
- Editorial interventions:
  - sic/corr, foreign

TEI P5 encoding: source description

```
<sourceDesc>
 <bibl>
  <title type="orig" xml:lang="sl-bohoric">Genovefa</title>
  <title type="reg" xml:lang="sl">Genovefa</title>
  <author>Schmid, Christoph von</author>
  <respStmt>
   <resp xml:lang="sl">Prevajalec</resp>
   <resp xml:lang="en">Translator</resp>
   <name>Malavašič, Fran</name>
  </respStmt>
  <date>1841</date>
  <publisher>J. Blasnik</publisher>
  <pubPlace>V Ljubljani</pubPlace>
  <extent>107</extent>
  <idno>10242 oder 13150 → NUK - Narodna in univerzitetna knjižnica</idno>
  <note type="TraDok" xml:lang="de">
  <ref target="http://itat2.uni-graz.at/pub/tradok/">TraDok</ref><lb/>
  1841<lb/>
```

TEI P5 encoding: text body

```
<pb n="[3]" facs="#FPG04260-002" xml:id="pb.003"/>
<div type="level1" xml:id="div.2">
  <head xml:id="head.2">1. <lb/>Kako Ožbalt iz vojske domú pride in kaj ljudjé govorijo.</head>
  <figure xml:id="figure.3">
    <figDesc>Ornamentna sličica. Okrašena črka v.</figDesc>
  </figure>
  <p xml:id="p.7">V nedéljo po poldne je bilo in v Zlati Vasi so mlajši fantini in dekleta pod staro lipo sedéli in peli, ali pa se smejali, kadar jo je kdó iz pivnice prilomil, ki je pregloboko v kozarček polukal. Nekteri kmetje s svojimi ženami so pa v gostivnici sedéli in pri bokalu prav židane volje bili, kakor je že to navada, kadar sta vino in vol po ceni.</p>
  <p xml:id="p.8">Kar jo primaha nék neznan človek v vas. Terdne in velike postave je bil in
   <choice>
     <sic>kakil</sic>
     <corr>kakih</corr>
   </choice>
tridesét lét je mogel iméti; obléčen je bil v sivi suknji, na strani je imel veliko sabljo, na herbtu pa…
```

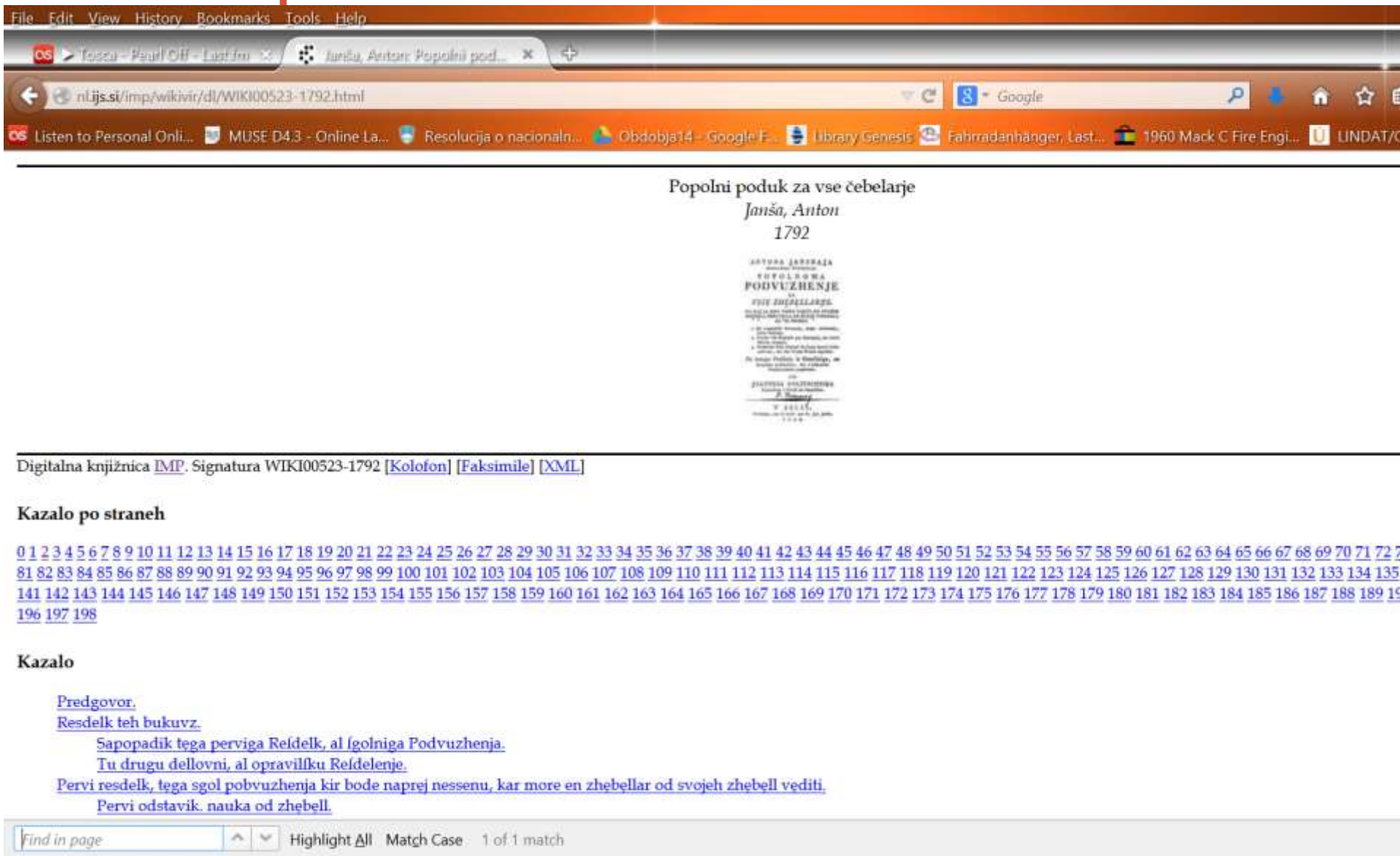
Up-translation to TEI

- AHLib:
  Dedicated RTF-to-TEI web converter
- IMPACT:
  XSLT stylesheets for PageXML-to-TEI conversion
- Wikisource:
  PHP converter for HTML-to-TEI
- Latest: http://nl.ijs.si/tei/convert/

Down-translation to HTML

- We use (slightly modified) TEI XSLT stylesheets for TEI-to-HTML
- Each text is one HTML file, showing both the facsimile and (typeset) transcription
- Indexes to books are by taxonomy, sorted by (one of)
  - author, title, date, signature (one index with title pages)

Example book: front



Example book: body

[108]

alge (Vaucheria), v kterem se ravno plasma (b) nabira, ki se razvija v roječo trosko, ki pozneje izstopi (c). P. 167. nam kaže ravno tisti

dogodek; roječe troske gredó namreč iz niti neke glive (Saprolegnia).

129 Razun v poprejšnjem popisane množitve in razploditve kryptogamov se nahajajo sosebno pri algah, mahovih in praprotih, dogodki, ki so podobni oploditvi, kakoršna se godi pri višjih rastlinah; naredi se namreč še le vsled tega, da se srečate in združite dve tvorini, stanica, zmožna daljne razvitve.

Pod. 168.
Enkalična alga

Pod. 169.
Nitasta gliva: 60krat povečan.

Najenovitnejša primera teh dogodkov sta *konjugacija* (pod. 168.), ki obstoji v tem, da se dve enostanični algi združite ter v eno edino plodilno stanico sorastete — in pa *kopulacija*, ki se godi pri nitastih glivah, kjer se dve stanici skup ležeta (pod. 169. I), se napihnete (II. do V.) ter razvijete v ene edino *ižesno trosko ali zigosporo* s, iz ktere izraste potem nova rastlina.

Med tem, ko v povedanih primerih delajo plodilni deli, ki so si v vsem popolnoma enaki, ima mnogo alg in vse višje tajnocvetke organe, ki se dadó primerjati s prašnicami ali antherami očitno cvetočih rastlin. Ti organi se zato tudi imenujejo *pelodke* (antheridije). V njih se razvijo kot plodilna telesca tako

Real example…

alge (*Vaucheria*), v kterem se ravno plasma (b) nabira, ki se razvija v roječo trosko, ki pozneje izstopi (c). P. [108] 167. nam kaže ravno tisti

dogodek; roječe troske gredó namreč iz niti neke glive (*Saprolegnia*).

129 Razun v poprejšnjem popisane množitve in razploditve kryptogamov se nahajajo sosebno pri algah, mahovih in praprotih, dogodki, ki so podobni oploditvi, kakoršna se godi pri višjih rastlinah; naredi se namreč še le vsled tega, da se srečate in združite dve tvorini, stanica, zmožna daljne razvitve.

Slika 84. Pod. 168.
Enkalična alga

Slika 85. Pod. 169.
Nitasta gliva: 60krat povečan.

Najenovitnejša primera teh dogodkov sta *konjugacija* (pod. 168.), ki obstoji v tem, da se dve enostanični algi združite ter v eno edino plodilno stanico sorastete — in pa *kopulacija*, ki se godi pri nitastih glivah, kjer se dve stanici skup ležeta (pod. 169. I), se napihnete (II. do V.) ter razvijete v ene edino *ižesno trosko ali zigosporo* s, iz ktere izraste potem nova rastlina.

Med tem, ko v povedanih primerih delajo plodilni deli, ki so si v vsem popolnoma enaki, ima mnogo alg in vse višje tajnocvetke organe, ki se dadó primerjati s prašnicami ali antherami očitno cvetočih rastlin. Ti organi se zato tudi imenujejo *pelodke* (*antheridije*). V njih se razvijo kot

Example index

Digitalna knjižnica IMP - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Digitalna knjižnica IMP

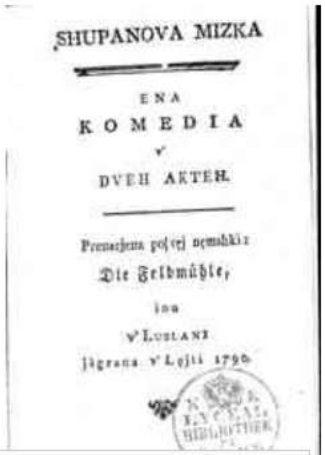
nl.ijs.si/imp/dl/index-facs.html

Google Koledar | The 51st Annual Meeti... | MySQL database on Li... | Git - Recording Chang... | Funding > Latest Info ... | SkE/Config/FullDoc – ... | Editorial Manager® | The Action | ISO Standards Develop... | tomaz erjavec

KAZALO

- Leposlovje
- Stvarna besedila
  - Knjige
  - Revije
  - Časopisi
- Verska besedila

Leposlovje


ŠHUPANOVA MIZKA

ENA

KOMEDIA

V

DVEH AKTEH.

Prestavjena poſrej nemshkiz

Die Feldmühle,

ino

V LUBLANI

Jgrana v' Lejti 1796


Leopolda Volkmera,

Fabule ino Péſmi.

Volkmer'vim shivlenjom

V Gradcu.

1826.


GENOVEFA.

Povéſt is ſtarih zhaſov

vſe dóbre ljudi,

matere in otróke.

KRISHTOFA ſHMIDA,

V LJUBLJANI.


VIII.

Martinez iz Podluna.


ĆUJTE, ĆUJTE

KAJ ŽGANJE DELA!

Prigodba

šalostna ino vesela za Slovence.

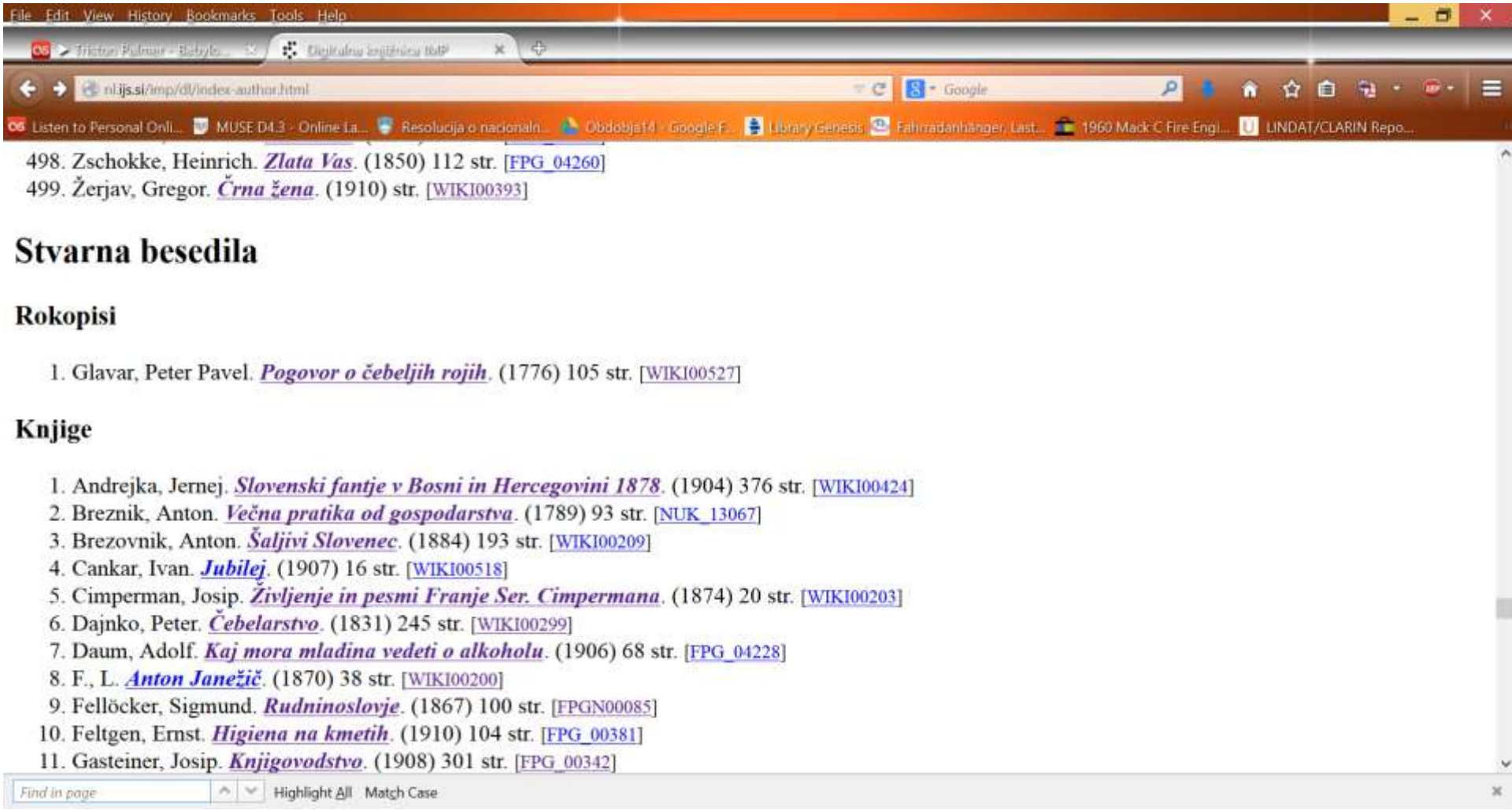
Poslovenil

F. . . Gl. . .

nl.ijs.si/imp/wikivir/dl/WIKI00375-1847.html

Find: Match case

Example index



IMP DL: summary

- Large library of Slovene historical texts
- Focus on the 19th and early 20th century
- Rich (and somewhat varied) TEI encoding
- HTML rendering, facsimile PDF, TEI
- Basic indexes, no search (poor man‘s DL)
- No ePub, PDF of transcription
- Still too small
- But is JSI the place to host a large & professionaly curated cultural heritage library?

Part II. Linguistic annotation

ljubesen	ljubezen	ljubezen
ljubésen	ljubezen	ljubezen
ljubesin	ljubezen	ljubezen
ljubésin	ljubezen	ljubezen
ljubezen	ljubezen	ljubezen
ljubézen	ljubezen	ljubezen
ljubezin	ljubezen	ljubezen
ljubézin	ljubezen	ljubezen
lubesan	ljubezen	ljubezen
lubesen	ljubezen	ljubezen
lubésen	ljubezen	ljubezen
lubęsen	ljubezen	ljubezen
lubesèn	ljubezen	ljubezen
lubésèn	ljubezen	ljubezen
lubęsèn	ljubezen	ljubezen
lubesęn	ljubezen	ljubezen
lubesn	ljubezen	ljubezen
lubésn	ljubezen	ljubezen
lubęsn	ljubezen	ljubezen
lubeſn	ljubezen	ljubezen
lubiesen	ljubezen	ljubezen

lubiesn	ljubezen	ljubezen
ljubeznih	ljubeznih	ljubezen
ljubesnijo	ljubeznijo	ljubezen
ljubésnijo	ljubeznijo	ljubezen
ljubeznijo	ljubeznijo	ljubezen
ljubéznijo	ljubeznijo	ljubezen
lubesnio	ljubeznijo	ljubezen
lubęsnio	ljubeznijo	ljubezen
ljubezníjo	ljubezníjo	ljubezen
ljubesni	ljubezni	ljubezen
ljubésni	ljubezni	ljubezen
ljubesní	ljubezni	ljubezen
ljubezni	ljubezni	ljubezen
ljubézni	ljubezni	ljubezen
lubesne	ljubezni	ljubezen
lubeſne	ljubezni	ljubezen
lubèſne	ljubezni	ljubezen
lubesni	ljubezni	ljubezen
lubésni	ljubezni	ljubezen
lubęsni	ljubezni	ljubezen
lubeſni	ljubezni	ljubezen

The goo300k corpus

Gold-standard linguistically annotated corpus of historical Slovene

- 300.000 words, 1.100 pages
- page-sampled from IMP DL
- manually annotated
- TEI encoded

Word-level annotations:

- Tokenisation
- Contemporary word-form
- Contemporary lemma
- Part-of-speech
- Gloss for archaic words

goo300k text distribution

	Texts		Pages		Words	
	89	100.0%	1,100	100.0%	290,587	100.0%
1584	1	1.1%	8	0.7%	5,794	2.0%
1695	1	1.1%	27	2.5%	8,519	2.9%
1768-1799	8	9.0%	155	14.1%	22,216	7.6%
1800-1849	17	19.1%	280	25.5%	104,428	35.9%
1850-1899	62	69.7%	630	57.3%	148,413	51.1%
book	78	87.6%	994	90.4%	191,411	65.9%
newspaper	11	12.4%	106	9.6%	97,959	33.7%
religious	22	24.7%	326	29.6%	63,926	22.0%
nonfiction	28	31.5%	310	28.2%	146,855	50.5%
play	11	12.4%	145	13.2%	20,308	7.0%
poetry	2	2.2%	38	3.5%	4,875	1.7%
prose	26	29.2%	281	25.5%	53,406	18.4%
original	22	24.7%	283	25.7%	125,980	43.4%
translation	67	75.3%	817	74.3%	163,390	56.2%

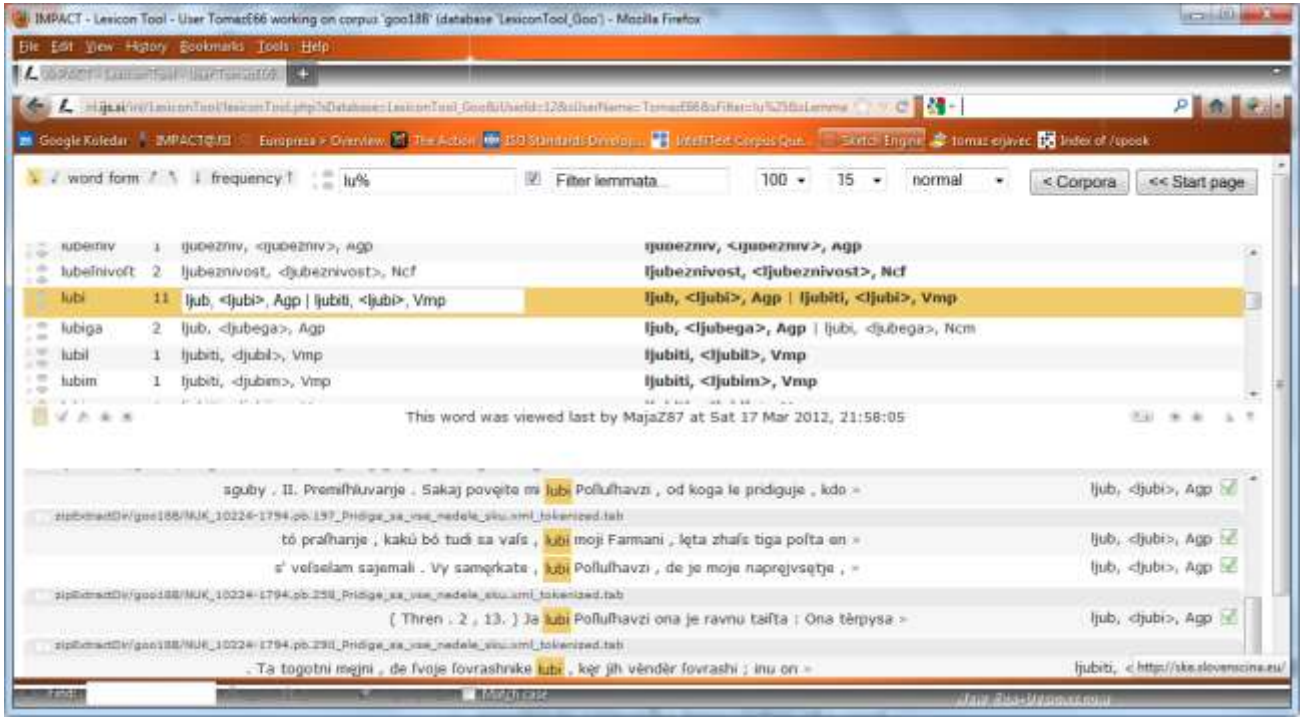
Linguistic annotation

bi	teiſte	isganjali	,	inu	vſe	shlaht	**Beteshe**	,	inu	vſe	shlaht	Bolesni	, osdraulali
bi	taiste	izganjali	,	in	vse	žlaht	**beteže**	,	in	vse	žlaht	bolezni	, ozdravljali
biti	taisti	izganjati	,	in	ves	žlaht	**betež**	,	in	ves	žlaht	bolezen	, ozdravljati
Va	P	Vmp	- C	P	Agp		Ncm	- C	P	Agp		Ncf	- Vmp
Gp	Z	Ggn	- V	Z	Ppn		**Som**	- V	Z	Ppn		Soz	- Ggn
-	-	-	- -	-	vsakovrsten		**bolečina**	- -	-	vsakovrsten		-	- -

```
<w lemma="biti" ana="#Va">je</w><c> </c>
<w lemma="pripravljati" ana="#Vmp">pripravljal</w><c> </c>
<w lemma="svoj" ana="#P">svoje</w><c> </c>
<choice>
  <orig><w>spremljevavce</w></orig>
  <reg><w lemma="spremljevalec" ana="#Ncm">spremljevalce</w></reg>
</choice><c> </c>
<w lemma="na" ana="#S">na</w><c> </c>
<w lemma="žalosten" ana="#Agp">žalostno</w><c> </c>
<choice>
  <orig><w>iznenadenje</w></orig>
  <reg><w lemma="iznenadenje" ana="#Ncn">iznenadenje</w>
    <desc><gloss>presenečenje</gloss><bibl>SSKJ</bibl></desc>
  </reg>
</choice>
<pc>,</pc>
```

Manual annotation

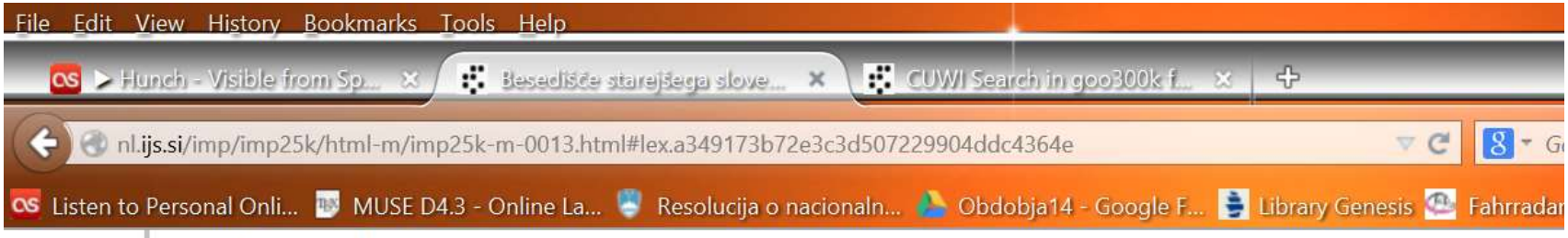
- Team of students lead by a lexicographer
- Several rounds of annotation and correction
- Annotators' manual, Cookbook, FAQ,…
- Used INL CoBaLT editor



IMP lexicon

- Simply a dump of the hand annotated corpus
- However, goo300k is too small for a (large) lexicon:
  - sampled another corpus from IMP DL, foo3M
  - also annotated in CoBaLT
  - but only word-forms that do not appear in goo300k
- The lexicon is (of course) also encoded in TEI P5
- Mounted in the Web as a set of HTML files (bespoke XSLT stylesheet)

Hyperlinked lexicon on the Web



▶ **izmed** predlog ↘ goo300k, IMP ↗ SSKJ, Pleteršnik

- izmed
  - is

    Svetu pismu noviga testamenta (1784): satoshili. 11. On pak je djal k' njim: Kdó bò *is* vaſs ta z
  - ismed

    Genovefa (1841): vſim ſvôjim velizhaſtvu ni bil tako lépo oblézhen, kakor êna

    Kmetijske in rokodelske novice (1843): in kakor bi jih boshjaſt svila, poginejo. — Nar lepſhi glava nar pervi

    Kmetijske in rokodelske novice (1843): s tako shlahtnimi ſerzi imenovati. S hvaleshnoſtjo ſe ſhe ma ſpomni, kakor patra Gabriela

    Kmetijske in rokodelske novice (1843): bilo, de ſo goſpod fajmoſhter, ki ſo porozhali, ſin eniga *isme* Krajnſkim
  - is med (združena oblika)

    Kmetijske in rokodelske novice (1843): trumo ſhe bolj divjih vojſhakov , ktere — kér ſo bili eni pſaglávze

Size of the lexicon

	XL		L		M		S	
Entries	28,096		21,653		12,255		4,156	
Lemmas	27,139	1.04	20,917	1.04	11,989	1.02	4,048	1.03
Modern word-forms	56,650	2.09	49,212	2.35	24,138	2.01	6,140	1.52
Historical word-forms	77,433	1.37	69,326	1.41	34,357	1.42	6,910	1.13
Five-tuples	81,740	2.91	73,263	3.38	35,941	2.93	7,021	1.69

- **XL**: all words
  (goo300k is fully annotated!)
- **L**: dictionary words
  (no numerals, typos, foreign words, proper nouns)
- **M**: historical words
  (historical form ≠ contemporary form)
- **S**: archaic words
  (with gloss)

ToTrTaLe

- Tool to automatically:
  1. Tokenise the text: mlToken
  2. Transcribe (modernise) the words: Vaam
  3. PoS tag the words: TnT
  4. Lemmatise the words: CLOG
- Use of background resources:
  - IMP lexicon + modern lexicon + transcription rules
  - tagger + lemmatiser models for contemporary Slovene
- TEI I/O

ToTrTaLe example

„… združimo tiga[tega] kužniga[kužnega]...“

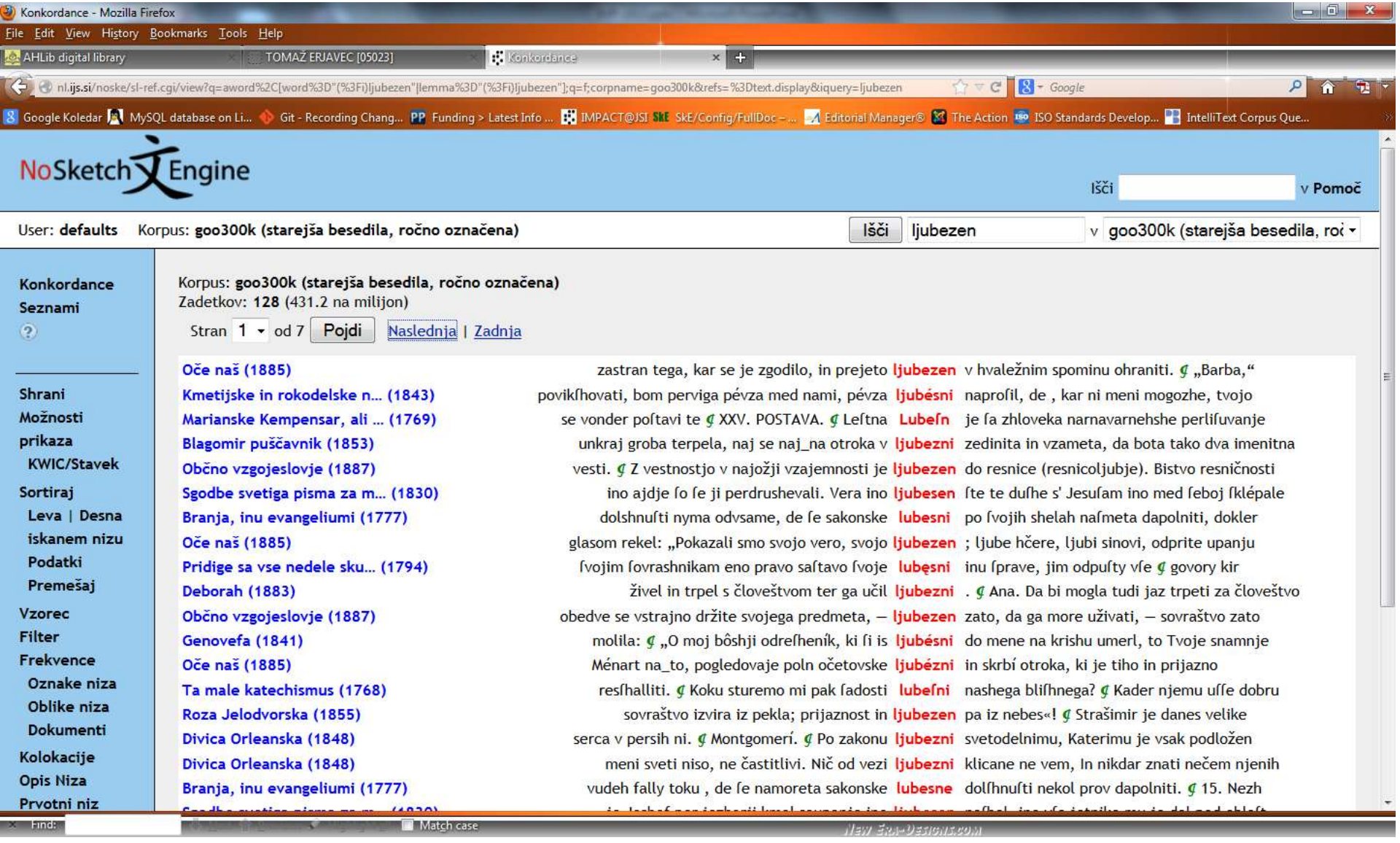
```
<w lemma="združiti" ana="#Vmer1p">združimo</w><c> </c>
<choice>
   <orig><w>tiga</w></orig>
   <reg><w lemma="ta" ana="#Pd-msg">tega</w></reg>
</choice>
 <c> </c>
 <choice>
   <orig><w>kužniga</w></orig>
   <reg type="pattern" n="[ega@←iga@]">
      <w lemma="kužen" ana="#Agpmsg">kužnega</w>
    </reg>
 </choice>
```

- Vmer1p = Verb, Type=main, Aspect=perfective, VForm=present, Person=first, Number=plural

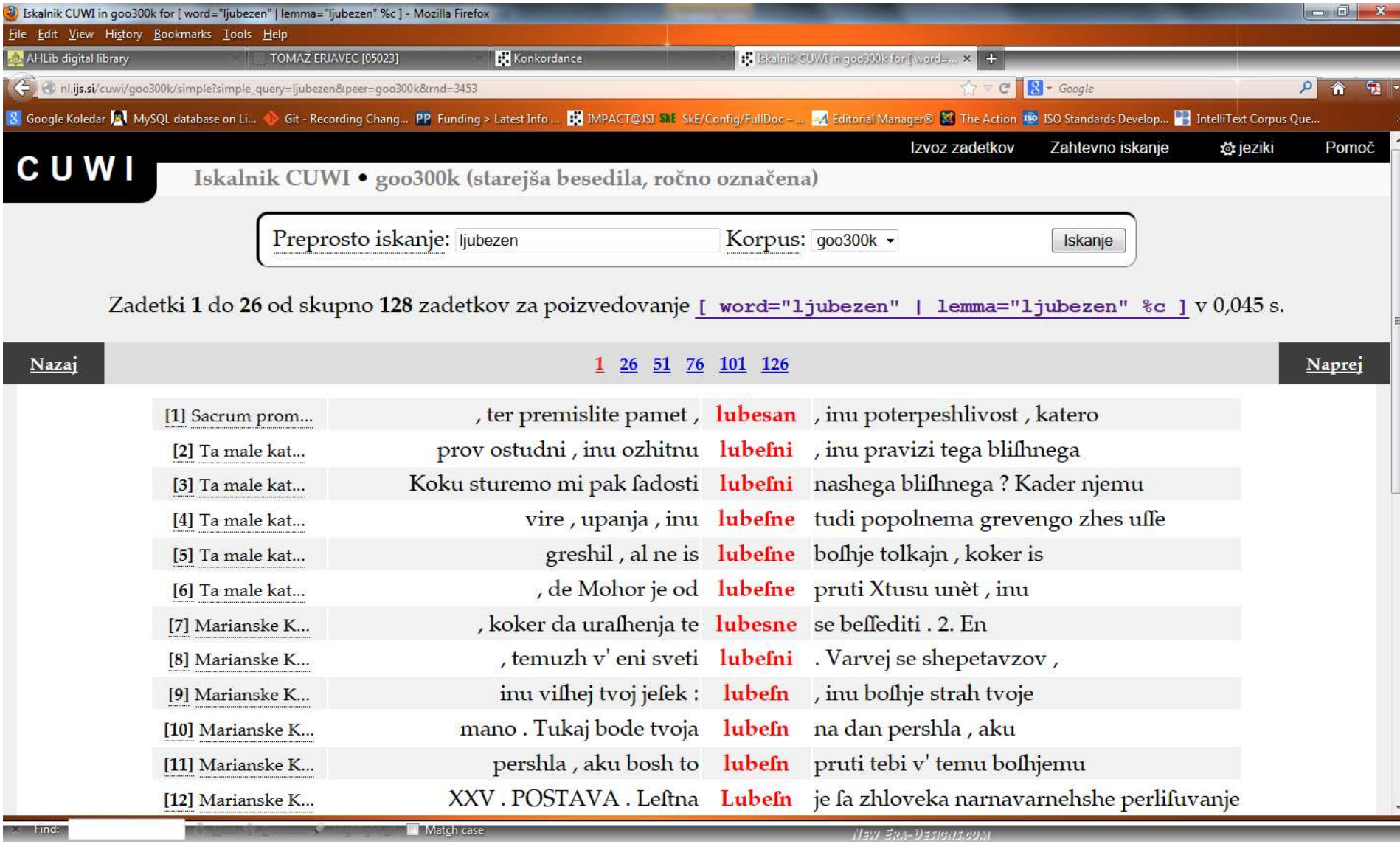
Concordancers

- The IMP DL was annotated with ToTrTaLe → IMP corpus
- The IMP and goo300k corpora are available via two concordancers:
  - **noSketchEngine**: OS version of the popular (and commercial) SketchEngine
  - **CUWI**: our front-end to the well-known IMS CWB corpus work-bench
- The concordancers offer:
  - powerful search query syntax (REs over words and annotations)
  - filters over meta-data (text types, year of publication, author, …)
  - various sorting options over concordances
  - construction of frequency lexica
  - collocations
  - saving results
  - etc.

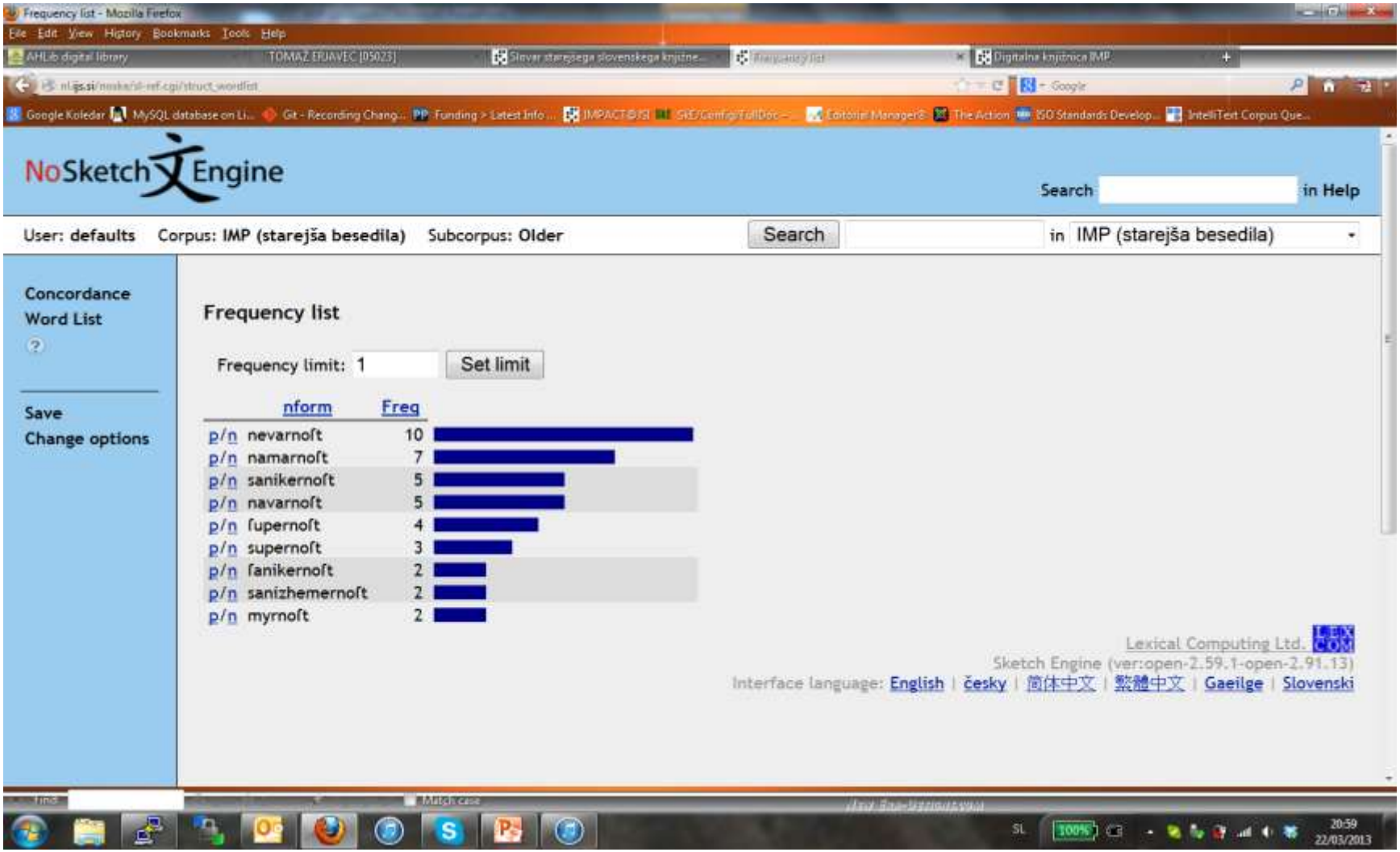
noSketchEngine



CUWI


Iskalnik CUWI in goo300k for [word="ljubezen" | lemma="ljubezen" %c] - Mozilla Firefox
File Edit View History Bookmarks Tools Help
AHLib digital library
TOMAŽ ERJAVEC [05023]
Konkordance
Iskalnik CUWI in goo300k for [word=...
nl.ijs.si/cuwi/goo300k/simple?simple_query=ljubezen&peer=goo300k&rnd=3453
Google
Google Koledar
MySQL database on Li...
Git - Recording Chang...
Funding > Latest Info ...
IMPACT@JSI
SkE/Config/FullDoc – ...
Editorial Manager®
The Action
ISO Standards Develop...
IntelliText Corpus Que...
CUWI
Izvoz zadetkov
Zahtevno iskanje
jeziki
Pomoč
Iskalnik CUWI • goo300k (starejša besedila, ročno označena)
Preprosto iskanje: ljubezen
Korpus: goo300k
Iskanje
Zadetki 1 do 26 od skupno 128 zadetkov za poizvedovanje [word="ljubezen" | lemma="ljubezen" %c] v 0,045 s.
Nazaj
1 26 51 76 101 126
Naprej
[1] Sacrum prom... , ter premislite pamet , lubesan , inu poterpeshlivost , katero
[2] Ta male kat... prov ostudni , inu ozhitnu lubeſni , inu pravizi tega bliſhnega
[3] Ta male kat... Koku sturemo mi pak ſadosti lubeſni nashega bliſhnega ? Kader njemu
[4] Ta male kat... vire , upanja , inu lubeſne tudi popolnema grevengo zhes uſſe
[5] Ta male kat... greshil , al ne is lubeſne boſhje tolkajn , koker is
[6] Ta male kat... , de Mohor je od lubeſne pruti Xtusu unèt , inu
[7] Marianske K... , koker da uraſhenja te lubesne se beſſediti . 2. En
[8] Marianske K... , temuzh v' eni sveti lubeſni . Varvej se shepetavzov ,
[9] Marianske K... inu viſhej tvoj jeſek : lubeſn , inu boſhje strah tvoje
[10] Marianske K... mano . Tukaj bode tvoja lubeſn na dan pershla , aku
[11] Marianske K... pershla , aku bosh to lubeſn pruti tebi v' temu boſhjemu
[12] Marianske K... XXV . POSTAVA . Leſtna Lubeſn je ſa zhloveka narnavarnehshe perliſuvanje
Find:
Match case

More than just concordances



Conclusions

- Presented IMP language resources, the result of three connected projects
- Available at http://nl.ijs.si/imp/
- A varied and connected BLARK for historical Slovene
- Used and useful for language technologies and digital humanities

HLT vs DH

- **Human language technologies:**
  digital library, lexicon, corpus
  - uniformly encoded in TEI P5
  - available under CC-BY
- **Digital humanities:**
  - web mounted digital library and lexicon
  - goo300k and IMP corpus available through concordancers
- Two birds with one stone (or is it two stones?)

HLT R&D

- Lexicon used for full-text search in dLib.si, the digital library of Slovenia
- Experiments investigating better modernisation, e.g.: Scherrer Y., Erjavec T. (2013) Modernizing historical Slovene words with character based SMT. In: *BSNLP 2013 - 4th Biennial Workshop on Balto-Slavic Natural Language Processing*, Sofia.

Use in digital humanities

- Server logs from 2013 and first half of 2014 show 2,000–3,400 accesses per month or about 100 per day
- On-line survey, 52 respondents:
  - 2/3: elementary and secondary school teachers, next under- and postgraduate students
  - 2/3: resources used about once a month
  - 90%: resources are useful for linguistic research, for teaching Slovene and important for the Slovene society
  - 95%: resources are worthy of recommendation
  - improvements should be focused on:
    linguists (14%), pupils and students (both 13%), elementary and secondary school teachers (both 12%) and lexicographers (9%)

The horror of cycles

- Changing annotation guidelines
- Automatic / manual annotation:
  ToTrTaLe → Corpus → CoBaLT → Lexicon → ToTrTaLe
- Text corrections:
  Corpus → CoBaLT☹ → a tangle of XSLT → Corpus

Further work

- HLT experiments with text normalisation
- ToTrTaLe Version 2
- Correcting mistakes
- Extend DL & corpus > 1918 (Wikisource)
- Offer more output formats: ePub, PDF
- Import into the „big“ DLs and dictionary portals