

# Next steps in newspaper digitization: making use of digitized texts at NLL

IFLA Newspapers pre-conference  
Geneva, 13.08.2014.

**Arturs Zogla**

Head of Digital Library  
National Library of Latvia  
[arturs.zogla@lnb.lv](mailto:arturs.zogla@lnb.lv)

# Newspaper digitization projects at NLL

- **2000-2006:**
  - Image-only PDFs (no OCR)
- **2007-2009:**
  - OCR with article-level segmentation
  - 40 titles/350 000 pages
- **2010-2012:**
  - Mass digitization with OCR and segmentation
  - ~3 million pages (1760-2011)

# Heritage-1

Laikraksts "Dienas Lapa", 1886. - 1905., 1918. gads			
Gadi:	1889. gads (Nr./datums)		
1886	<u>1/2.01.</u>	<u>2/3.01.</u>	<u>3/4.01.</u>
1887	<u>6/9.01.</u>	<u>7/10.01.</u>	<u>8/11.01.</u>
1888	<u>11/14.01.</u>	<u>12/16.01.</u>	<u>13/17.01.</u>
1889	<u>16/20.01.</u>	<u>17/21.01.</u>	<u>18/23.01.</u>
1890	<u>21/26.01.</u>	<u>22/27.01.</u>	<u>23/28.01.</u>
1891	<u>26/1.02.</u>	<u>27/2.02.</u>	<u>28/3.02.</u>
1892	<u>31/7.02.</u>	<u>32/8.02.</u>	<u>33/9.02.</u>
1893	<u>36/13.02.</u>	<u>37/14.02.</u>	<u>38/15.02.</u>
1894	<u>41/18.02.</u>	<u>42/20.02.</u>	<u>43/21.02.</u>
1895	<u>46/24.02.</u>	<u>47/25.02.</u>	<u>48/27.02.</u>
1896	<u>51/4.03.</u>	<u>52/6.03.</u>	<u>53/7.03.</u>
1897	<u>56/10.03.</u>	<u>57/11.03.</u>	<u>58/13.03.</u>
1898	<u>61/16.03.</u>	<u>62/17.03.</u>	<u>63/18.03.</u>
1899	<u>66/22.03.</u>	<u>67/23.03.</u>	<u>68/24.03.</u>
1900	<u>71/28.03.</u>	<u>72/29.03.</u>	<u>73/30.03.</u>
1901	<u>76/3.04.</u>	<u>77/4.04.</u>	<u>78/5.04.</u>
1902	<u>81/11.04.</u>	<u>82/12.04.</u>	<u>83/13.04.</u>
1903	<u>86/17.04.</u>	<u>87/18.04.</u>	<u>88/19.04.</u>
1904	<u>91/22.04.</u>	<u>92/24.04.</u>	<u>93/25.04.</u>
1905	<u>96/28.04.</u>	<u>97/29.04.</u>	<u>98/1.05.</u>
1918	<u>101/4.05.</u>	<u>102/5.05.</u>	<u>103/8.05.</u>
	<u>106/11.05.</u>	<u>107/12.05.</u>	<u>108/13.05.</u>
	<u>111/19.05.</u>	<u>112/20.05.</u>	<u>113/22.05.</u>
			<u>4/5.01.</u>
			<u>9/12.01.</u>
			<u>14/18.01.</u>
			<u>19/24.01.</u>
			<u>24/30.01.</u>
			<u>29/4.02.</u>
			<u>34/10.02.</u>
			<u>39/16.02.</u>
			<u>44/22.02.</u>
			<u>49/28.02.</u>
			<u>54/8.03.</u>
			<u>59/14.03.</u>
			<u>64/20.03.</u>
			<u>69/25.03.</u>
			<u>74/31.03.</u>
			<u>79/6.04.</u>
			<u>84/14.04.</u>
			<u>89/20.04.</u>
			<u>94/26.04.</u>
			<u>99/2.05.</u>
			<u>104/9.05.</u>
			<u>109/16.05.</u>
			<u>114/23.05.</u>

[http://data.lnb.lv/digitala\\_biblioteka/laikraksti/](http://data.lnb.lv/digitala_biblioteka/laikraksti/)

# OCR pilot project [1]

- **Content:**
  - 40 titles/350 000 pages
  - Periodicals published: 1920-1940
  - Antiqua script/out of copyright
- ***Olive Software:***
  - OCR (without manual corrections) and article level segmentation
  - Website with *search, browse and advanced search*

# OCR pilot project [2]

Search articles:

Last results: [search](#) [read newspaper](#) [browse newspapers](#)

Rīts: Wednesday, April 10, 1940

[Download issue \(PDF\)](#) [Print](#) [Save article as HTML](#) [Send article by e-mail](#) [Text View](#) [Show Page \(16\)](#)

## Search results

You searched: **oslo** in publications: **Bauskas Vēstnesis, Brīvzeme, Brīvais Zemnieks, Cēsu Vēstis, Ciņa, Darbs, Darba Iekšlietu Ministrijas Vēstnesis, Izglītības Ministrijas Mēnešraksts, Jēkabpils Vēstnesis, Kurzemes Vārds, Latvijas Kareivis, Laikonis, Nacionālā Zemgale, Ogres Ziņas, Padomju Kuldīga, Padomju Ziņas, Rīgasche Rundschau, Rīts, Sporta Pasaule, Saeima, Vēstnesis, Ventas Balss, Zemgale, Zemgales Balss, Zemgales Vēstnesis** (results )

1 2 3 4 5 6 7 N

### Pirmā telefoniskā saruna ar Oslo pēc okupācijas

*Specialziņojums «Rītam».*

G. S t o k h o l m ā, 10. aprīlī (pl. 1.30). «United Press» pārstāvim Stokholmā šonakt bijusi telefoniska saruna ar Oslo pēc Norvēģijas galvas pilsētas okupācijas. Pateicoties sevišķai iestāžu labvēlībai, lielās amerikāņu agentūras «United Press» korespondents Oslo guvis atļauju sazināties ar savu Stokholmas kolegu un isumā atferēt savus iespaidus. Cik līdz šim zināms, tā ir vienīgā telefona saruna ar Oslo, ko pēdējās 24 stundās Norvēģijas galvas pilsētā guvusi privāta persona.

«United Press» korespond. Oslo stāsta, ka Norvēģijas galvas pilsētā valdot miers un ka pēc ciņu izbeigšanās iedzīvotājos nescot manāms satraukums. Tie arī neizrādot pretestību vācu karaspēkam, kas patrulē ielās. «United Press» korespondents redzējis 300—400 vācu karavīrus, kas viegliem ložmetēju un lielgabalu ratiem devušies cauri pilsētai. Vācu desants tā tad pirmām kārtām ved sev līdz zirgus, kā mehāniskie transporta līdzekļi. Oslo ielās arī redzēti vairāki automobiļi, kuros blakus vācu virsniekiem sēdējuši arī norveģu virsnieki. Tie, domājams, bijuši sagūtīti Oslo garnizona virsnieki. «United Press» korespondents tālāk vēl stāstīja par vērojumiem Norvēģijas galvas pilsētā un norveģu tautā. Saruna pārtraukta tajā brīdī, kad amerikāņu žurnālists minēja, ka pēc viņa rīcībā esošām ziņām arī vācu karakuģi nesot apšaudījuši pilsētu, tikai Ulevalas priekšpilsētā kritušas vairākas granātas, nodarot bojājumus arī dzīvojamām ēkām.

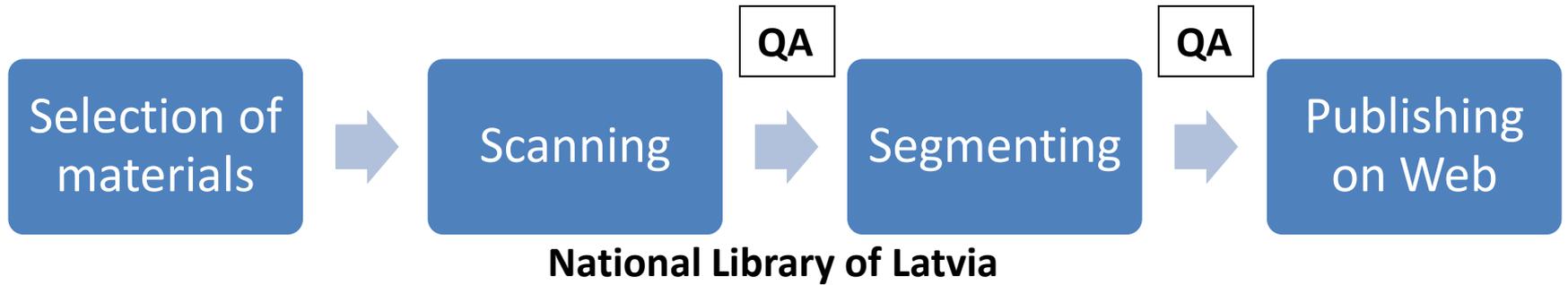
Drāmatiskos un neparastos apstākļos norisinājusies Oslo aerodroma ieņemšana. Laikā starp pl. 7 un 9 šorīt vācu lidmašīnas nolaikušās Oslo apkārtnē un izcēlušās gaisa desantas, kas ieņēmuši vairākas Oslo pilsētas daļas. Oslo aerodromā nolaidušies 30 smagie vācu bumbvedēji, kuriem pretim norveģi varējuši stāties tikai ar vienu pretaviācijas bateriju. Nometuši dažas bumbas un apšaudījuši aerodromu ar ložmetējiem, vācu bumbvedēji izcēlušī brupotus karavīrus, kas šā ciņā pārvarējuši aerodroma sardzi. Līdzīgā kārtā okupēti arī citi aerodromi Dienvidnorvēģijā.

Website no longer exists

# Mass-digitization

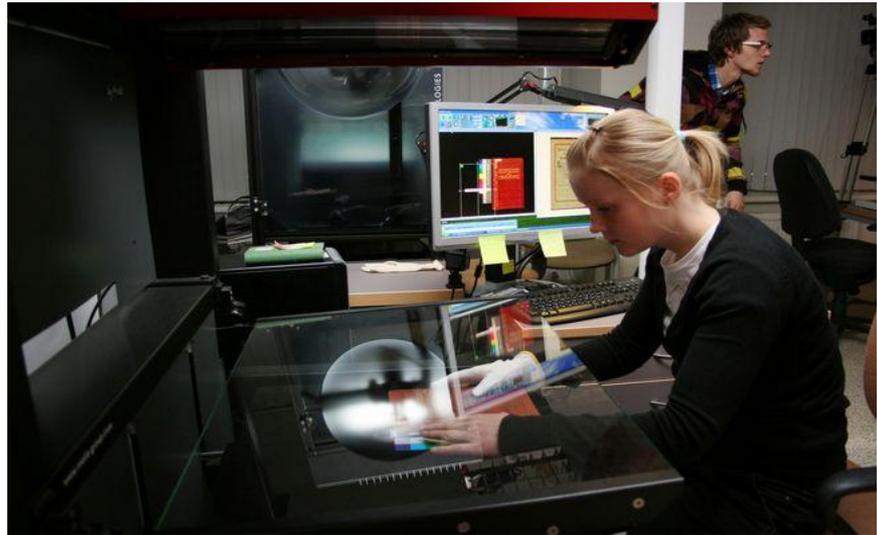
- Funded by *European Regional Development Fund*
- Time frame: 2010.-2012.
- Scope of the project:
  - Periodicals: ~3 million pages (~1000 titles)
  - Books: ~1.4 million pages (~7000 books)
  - Portal of searchable historical texts
  - User interactivity

# Digitization process



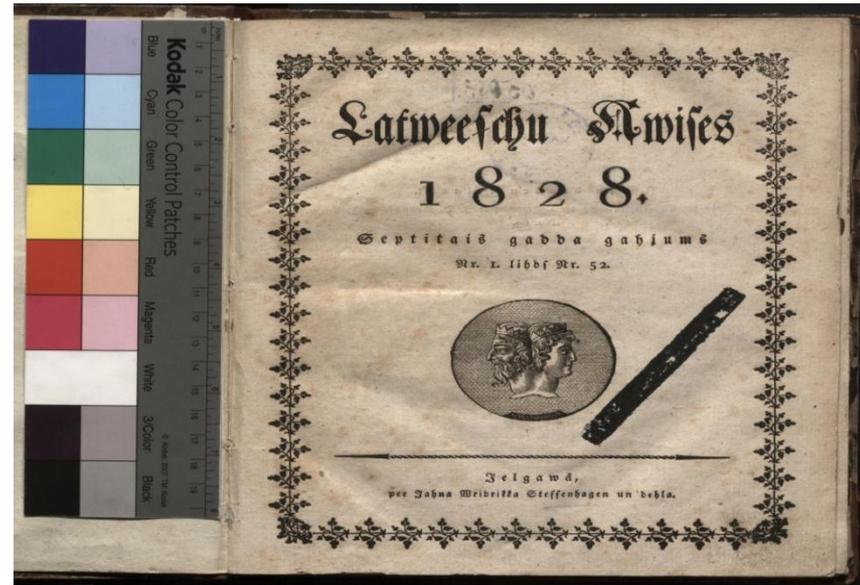
# Scanning

- Outsourced to  laitelecom BPO
- 2 shipments per month:
  - ~46 000 pages



# Scanning output

- JPEG 2000 file for each page
  - Books, magazines: RGB
  - Newspapers: Greyscale
  - Resolution: 400 dpi
- File size: 3-100 MB



# Segmentation

- Outsourced to LETA (*CCS docWorks*)
- **Zoning:**
  - Articles/titles
  - Images/captions
  - Authors
  - Tables
  - Advertisements
- **OCR:**
  - ABBYY Finereader engine
  - Manual correction of titles and image captions
  - For long titles/captions only Named Entities corrected

# Segmentation facts

- **Output:**
  - 1 METS file per issue
  - 1 PDF file per issue
  - 1 ALTO file per page
  - 1 JPEG file per page
  - 1 OCR file per article
- **Required work force:**
  - up to 60 operators
  - Three shifts

# OCR quality [1]

- **Antiqua script**
  - Most cases: 98-100%
  - 108% case where OCR engine corrected a typo in original text
- **Gothic script**
  - ~80-90% characters recognized correctly

# OCR quality [2]

Ne weenu ween reisi dsirdets par sihdainu samainischanu dsendeschanas eestahdēs. Lai no schahdsem gadijumeem iswairitos un nomeerinatu pahrbeedetās mahtes, modernajās dsendeschanas eestahdēs jaunpeedsimuscheem noņem kahjas nospeedumu. Schis nospeedums ir tāpat pagatavots, kā kriminalajā daktiloskopijā.

Tuhliņ pehz dsimschanas wezmahtes un ahrsta usraudsiņa, kas bijuschi klaht pee dsemdibam, isdara kahjas nospeedumu. Ahrsteem schi mahksla jāstude tāpat, kā kriminalpolizisteem.

Lai nebuhtu ne masako schaubu, tad blakus behrna kahju nospeedumam nospeesch ari mahtes pirkstu. Ja rodas pahrpratumi, tad nospeedumus nodod kriminalpolizijas daktiloskopiskajā nodaļā un tur galīgi pahrleezinajas, waj nospeedumi atbilst atteezīgai mahtei un behrnam.

Original

Ne meenu Meen reisi dsirdets par sihdainu samainischanu dsendeschanas eestahdss. Lai no schahdsem gadijumeem iswairitos un nomeerinatu pahrbeedetās mahtes, modernajās dslnndeschanas eestahdes jaunpeedsimuscheem noņem kahjas nospeedumu. Schis nospeedums ir tāpat pagatamots, ka kriminalajā daktiloskopijā.

Tuhliņ pehz dsimschanas mezmahhtes un ahrsta usraudsiņa, kas bijuschi klaht pee dsemdibam, isdara kahjas nospeedumu. Ahrsteem schi mahksla jāstude' tāpat, kti kriminalpolizisteem.

Lai nebuhtu ne masako schaubu, tad blakus behriui tahju nospeedumam nospeesch ari mahtes pirkstu, Ja rodas pahrpratumi, tad nospeedumus nodod kriminalpolizijas daktiloskopiskajā nodaļā un tur galīgi pahrleezinajas. maj nospeedumi atbilst atteezīgai mahtei un behrnam.

OCR

Correct characters: 685/739 (~92.7%)

Correct words: 49/90 (only ~54,4%)

# Portal

← → ↻ www.periodika.lv

BOOKS > PERIODICALS ▼ MAPS > PICTURES > AUDIO & VIDEO > WWW ARCHIVE > COLLECTIONS >

Login | LV | RU | EN

**L N B**  
DIGITĀLĀ BIBLIOTĒKA

**PERIODICALS**

My bookmarks | My recommendations | My collections | My corrections | My comments | My rated

Advanced search

**BROWSE**

Date: 1038

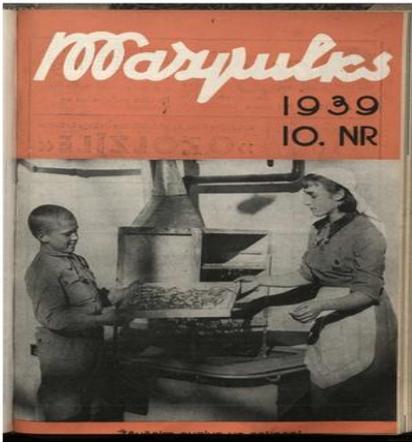
JAN	FEB	MAR
APR	MAY	JUN
JUL	AUG	SEP
OCT	NOV	DEC

MO	TU	WE	TH	FR	SA	SU
25	26	27	28	29	30	31
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31	1	2	3	4

**Periodicals:**

- 3. Vispasaules Latvieš...
- Adelaides Latviešu Zīp...
- Aistija: Latviešu liet...
- Aizdedze: 323. transpo...
- Aizkulise
- Aizkulises
- Aizputes Vēstnesis
- Akadēmiskā Dzīve
- AKKA Pakete

**Editor's choice**



Title: Mazpulks

**Lačplēšiem autobusus jāmaksā puscena**

Uz soseju un zemes ceļu dep-ta pārziņa esošām autobusu līnijām Lačplēša kara ordeņa kavalieriem jāmaksā puse no bilešu cenas, arī uz nupat Dzelzeju virsvaldes pārņemto Rīgas - Jelgavas autobusu līniju.

**Plostošanu traucē augstais ūdens līmenis**

Līdz vakardienai pa Daugavu līdz Rīgai bija nopludināti šīn sezonā 150 plosti. Dienā patlaban pienāk Jumpravmuižā 15—25 vietējo plostu. Plostošanu stipri traucē ārkārtīgi augstais ūdens līmenis.

Title: Rīts

# Portal

The screenshot shows a web-based digital viewer for a historical document. The browser address bar displays the URL: [www.periodika.lv/periodika2-viewer/view/Index-dev.html?lang=en#issue:/p\\_001\\_mazp1939n10|issueType:p](http://www.periodika.lv/periodika2-viewer/view/Index-dev.html?lang=en#issue:/p_001_mazp1939n10|issueType:p). The page title is "Mazpulks, Nr.10 (15.10.1939)".

On the left side, there is a navigation menu with a search bar and a list of contents:

- Quick Search
- Mazpulks, Nr.10 (15.10.1939)
  - Nevienam neatdosim mūsu zemi un mūsu rīcības brīvību
  - Zeme
  - Kas jauns mūsu zemē un valstī
  - AUGSIM LATVIJAI
  - Pļava
  - Divas nedēļas Zviedrijā
  - Centies saglabāt ziemā to, ko tu savas veselības nodrošināšanai atstāji pavasarī
  - Par kultūraugu slimību paraugu kārtošanu
  - Paparžu zemes stumbru vākšana
  - Sveču pagatavošana mājās apstākļos
  - Ziepju pagatavošana
  - Tas jāievēro mazpulku dalībniekiem
  - Ar mazpulku pateicības ziņi apbalvotie:
  - Par centību apbalvotie vadītāji
  - Mazpulku dzīve
    - Svētku dāvināšana
    - Latvijas mazpulku vadība Vācijā
    - 1939. g. iecelti mazpulku diženie
  - Pašu darbs
    - Ko es no tēva mācījos
  - Šahs
  - Pazīņojumi
  - Humors
  - Saturs
  - Section

The main content area displays six pages of the newspaper in a grid:

- Page 431: Cover of "Mazpulks" 1939, 10. NR, featuring an illustration of a man and a woman in a kitchen.
- Page 432: Advertisement for "OZOLZĪLE" (Oak Bark) and "Jelgavas audums" (Jelgava fabric), including contact information for K. A. Bušs and A. Ratfelders.
- Page 433: Article titled "Nevienam neatdosim mūsu zemi un mūsu rīcības brīvību" (We will not give our land and our freedom of action to anyone).
- Page 434: Textual content, likely an article or report.
- Page 435: Textual content, likely an article or report.
- Page 436: Textual content, likely an article or report.

At the bottom left, there is a "Rights" section with the status "Orphan".

Viewer based on a tool developed at **National Library of Luxembourg**

What next?

# Interactivity

- **OCR corrections**
- Comments
- Favourite articles
- Sharing on Twitter, Facebook, etc.
- Private collections

# OCR corrections [1]

The screenshot shows a web browser window with the address bar displaying the URL: [https://periodika.lndb.lv/periodika2-viewer/view/index-dev.html?lang=en#panel:pa|issue:/p\\_000\\_xrit1997n531|article:DIVL304|issueType:P](https://periodika.lndb.lv/periodika2-viewer/view/index-dev.html?lang=en#panel:pa|issue:/p_000_xrit1997n531|article:DIVL304|issueType:P). The browser title is "Ritums, Nr.531 (01.09.1997) - Google Chrome".

The main content area displays a news article with the following text:

## SWEDEN AND LATVIA ABOLISH VISAS

Sweden and Latvia on 4 August reached an agreement on visa-free travel. However, Sweden said the deal would not become valid until Latvia signed the Geneva convention on refugees without regional reservations. "Many countries have made certain amendments to, or interpretations of, the Geneva convention but we think Latvia has made too many reservations," a spokeswoman at the Swedish foreign ministry told Reuters. "In particular, Latvia only acknowledges refugees from a limited number of countries."

On the right side of the browser window, there is an "OCR edition" panel. It includes a "Save" button and a list of text lines for editing. The line containing the quote "Many countries have made certain amendments to, or interpretations of, the Geneva convention but we think Latvia has made too many reservations," is highlighted in blue, indicating it is the current selection for editing.

# OCR corrections [2]

Today This week This month All time

< Page 1 from 3 >

Found: 41 results Show on page: 20 40

1. Aigars Liepiņš (5678)
2. Raivis Kļaviņš (636)
3. Sven Lepa (469)
4. Gunita Kulmane (298)
5. Danute Dūra (209)
6. Antra Indriksone (137)
7. Kitija Mirončuka (84)
8. Antra Indriksone (68)
9. Uldis Zariņš (53)
10. Māris Štemers (50)
11. Antra Krauze (44)
12. Marika Karlsona (40)
13. Dagnija Avota (Tobisa) (35)
14. Uldis Žimants (33)
15. Lita Vēvere (19)
16. Armands Kūceņš (18)
17. Karina Bandere (17)
18. Anda Baklāne (9)
19. Baiba Logina (9)
20. Jānis Rozefelds (9)

< Page 1 from 3 >

Found: 41 results Show on page: 20 40

*TOP 20 OCR error correctors*

# OCR corrections [3]

- Reasons for low interest in OCR correction:
  1. Relatively high levels of OCR quality.
  2. Lack of publicity of OCR correction features.
  3. Requires registration.

Use of digitized newspapers by third parties

# “Baricadopedia” [1]

- Project that collects materials on Latvia’s struggle for independence in 1980ies.
- **NLL:**
  - provided newspapers
- **“Baricadopedia” team:**
  - Manually corrected all OCR errors
  - Tagged each article

# “Baricadopedia” [2]

## «Palikšu partijā»

Arvids Dauders, Viestarts Gailītis

«Latvijas Jaunatne», 1990. gada 13. janvāri, Nr. 9

**Intervija ar Lietuvas KP CK un organizācijas «Sajūdis» seima locekli ROMUALDU OZOLU.**

— Jūs reizē esat «Sajūdis» seima un LKP CK loceklis, tātad darbojaties divu kardināli atšķirīgu organizāciju vadībā. Vai tā nav pretruna?

— Redzu pretrunas visur un visā. Ja turamies tikai pie principa, ka KP ir tā organizācija, kas noziegusies pret Lietuvas tautu, tad mums būtu jāpakaras. Taču, ja domājam par Baltijas iedzīvotājiem, tad jādarbojas Lietuvas KP rindās, ņemot vērā, ka tā ir atdalījusies no PSKP. Tas tad būs reāls ieguldījums Lietuvas tautas labā, jo noliegt iepriekšējo iespējams tikai, pārejot jaunā kvalitātē, proti, jaunā kompartijā. Tā ir alternatīva tai kompartijai, kas iznīcināja Lietuvas Republiku. Mans uzdevums ir, piederot pie organizācijas «Sajūdis» un reizē atrodoties KP (būtībā tā ir noziedzīga), piespiest to atteikties no savas pagātnes un likvidēt noziedzības sekas. Saukt pie atbildības komunistus noziedzniekus.

— **Pieņemsim, ka no «Sajūdis» tiek izslēgti visi tā rindās esošie komunisti. Kur tad jūs paliksiet?**

— Es palikšu partijā.

— **Iemesli?**

— Manuprāt, nav cita spēka, kas reāli varētu atrisināt Lietuvas neatkarības



Avots:	<u>Latvijas Jaunatne</u>
Nr.:	<u>9</u>
Datums:	<u>13.01.1990</u>
Tirāža:	<u>226142</u>
Lappuse:	<u>3</u>
Virsraksts:	<u>«Palikšu partijā»</u>
Rubrika:	<u>Tikšanās Lietuvā</u>
Tēmas:	<u>Lietuvas ekonomiskā blokāde,</u> <u>1990</u> <u>Lietuvas neatkarības cīņas</u>
Nozares:	<u>Politika</u>
Organizācijas:	<u>«Sajūdis»</u>
Mediji:	
Civēki:	<u>Romualds Ozols</u>
Vietas:	<u>Lietuva</u>
Notikums:	<u>E1990011100</u>
Gads:	<u>1990</u>

# Twitter



The correct spelling of “pineapple” in Latvian. Is it:

**ANANASS** or **ANANĀSS?**

137 mentions in  
[periodika.lv](http://periodika.lv)

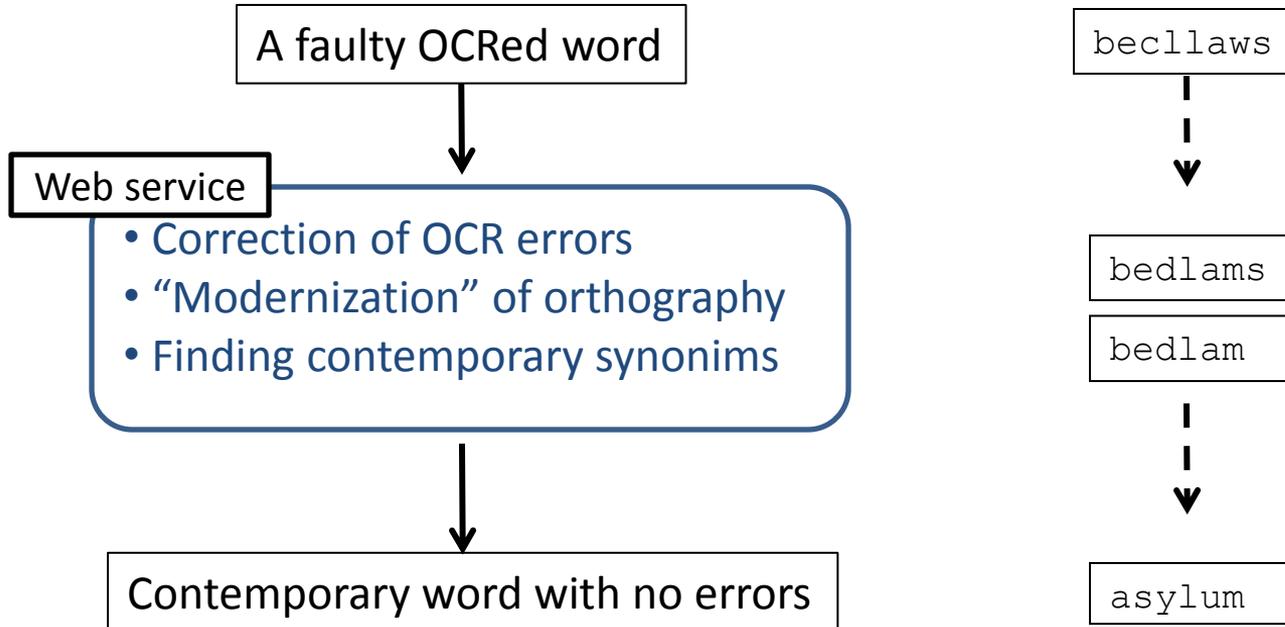
32 mentions in  
[periodika.lv](http://periodika.lv)

Old word “modernization” service

# A problem

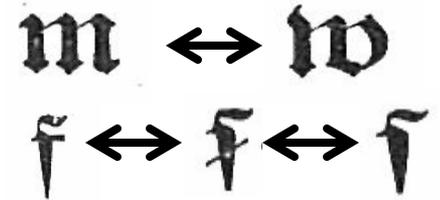
- Gothic texts
  - contain a lot of OCR errors;
  - are written in obsolete orthography;
  - contain obsolete words.
- Low hopes for crowd-sourcing

# Solution



# Transliteration rules

Rule	Type
m → w	OCR error / optional
f → s	OCR error / optional
w → v	Orthography / mandatory
ah → ā	Orthography / mandatory
tsch → č	Orthography / mandatory
ee → ē	Orthography / context-sensitive
ee → ie	
ee → ee	



# Results

```
remoluzionara [Orģināls] -> remoluzionara [Translit.] -> remoluzionars [Ar minēšanu] -> revolucionārs LLVW [Aptuveni]
remoluzionara [Orģināls] -> remoluzionara [Translit.] -> remoluzionarš [Ar minēšanu]
remoluzionara [Orģināls] -> remoluzionara [Translit.] -> remoluzionaris [Ar minēšanu]
remoluzionara [Orģināls] -> remoluzionara [Translit.] -> remoluzionara [Ar minēšanu]
remoluzionara [Orģināls] -> remoluzionara [Translit.] -> remoluzionart [Ar minēšanu]
```

Generated word  
variants

Word is found in  
a dictionary

- On average:
  - ✓ 2.89 word variants for an input word
  - ✓ 92.45% probability that the correct one is among them

# Three computer linguistics experiments

# Three experiments

1. Newspaper text corpus
2. Named Entity Recognition
3. Time-sensitive dictionaries

# Newspaper text corpus

- Advanced indexing of newspaper text
- Advanced full-text search query language
- 4.5 billion tokens = largest public text corpus for Latvian language

# Named Entity Recognition [1]

- 7 types of tags:
  - Person
  - Location
  - Organization
  - Facility
  - Event
  - Product
  - Time
- 21 subtypes

# Named Entity Recognition [2]

- Stanford CFR classifier
- Ground truth: manually tagged sample documents with ~150 000 words
- Thesauri used:
  - Persons
  - Places
  - Institutions



*UN = United Nations*

# Named Entity Recognition [3]

- How to tag?

"President of Republic of Latvia Guntis Ulmanis met with Minister of Culture today."

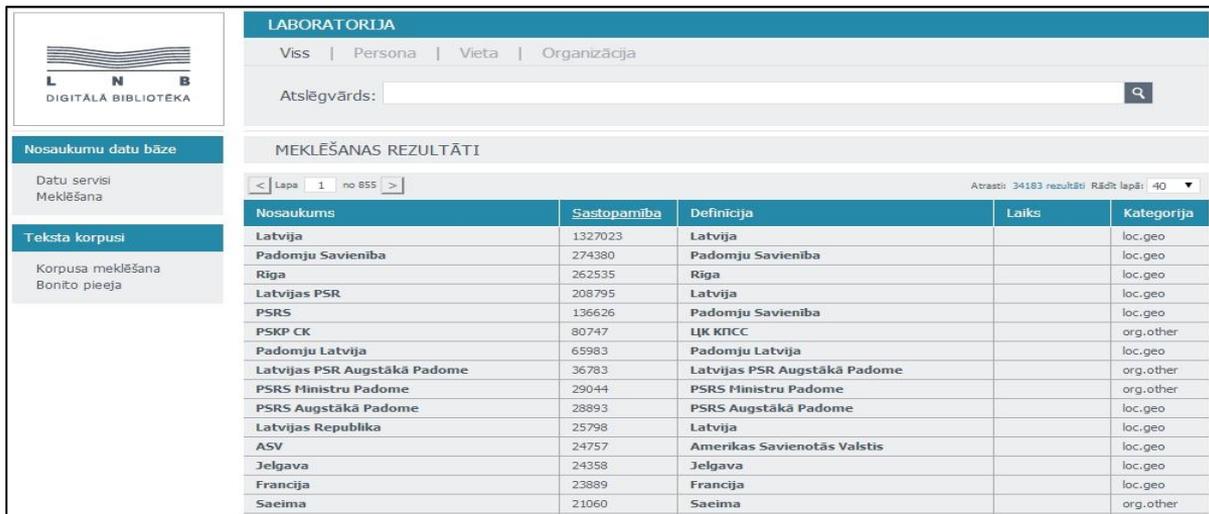
"**President of Republic of Latvia Guntis Ulmanis** met with **Minister of Culture** today."

"President of **Republic of Latvia** **Guntis Ulmanis** met with **Minister of Culture** today."

"President of **Republic of Latvia** **Guntis Ulmanis** met with **Minister of Culture <of Republic of Latvia>** today."

# Named Entity Recognition [4]

- Result
  - 26 000 entities that were found at least 10 times
  - >80% for good quality OCR texts



The screenshot shows the search results page for 'LABORATORIJA' on the National Library of Latvia (LNB) website. The page includes a search bar, navigation tabs, and a table of results.

**LABORATORIJA**

Viss | Persona | Vieta | Organizācija

Atslēgvārds:

**MEKLĒŠANAS REZULTĀTI**

< Lapa 1 no 855 >

Atrasti: 34183 rezultāti Rādīt lapas: 40

Nosaukums	Sastopamība	Definīcija	Laiks	Kategorija
Latvija	1327023	Latvija		loc.geo
Padomju Savienība	274380	Padomju Savienība		loc.geo
Rīga	262535	Rīga		loc.geo
Latvijas PSR	208795	Latvija		loc.geo
PSRS	136626	Padomju Savienība		loc.geo
PSKP CK	80747	ЦК КПСС		org.other
Padomju Latvija	65983	Padomju Latvija		loc.geo
Latvijas PSR Augstākā Padome	36783	Latvijas PSR Augstākā Padome		org.other
PSRS Ministru Padome	29044	PSRS Ministru Padome		org.other
PSRS Augstākā Padome	28893	PSRS Augstākā Padome		loc.geo
Latvijas Republika	25798	Latvija		loc.geo
ASV	24757	Amerikas Savienotās Valstis		loc.geo
Jelgava	24358	Jelgava		loc.geo
Francija	23889	Francija		loc.geo
Saeima	21060	Saeima		org.other

# Time-sensitive dictionaries [1]

- «President of USA»
  - *J. F. Kennedy* in 1959 article
  - *Barack Obama* in 2014 article
- «Bombay» in 1970ies = «Mumbai» today
- 1 dollar in 1933  $\neq$  1 dollar in 2014
- Street names in a city change over time

# Time-sensitive dictionaries [2]



What address from 1930ies advertisment is this today?

# Time-sensitive dictionaries [3]

- Web service that
  - Converts old NE to current versions
  - Converts modern NE to a version for a particular year
  - Interprets keywords (like, «*president of USA*») for a particular year

# Thank you!

Arturs Zogla

[arturs.zogla@lnb.lv](mailto:arturs.zogla@lnb.lv)