Digitization of Historical Texts at the National Library of Latvia

Arturs ZOGLA, Jurgis SKILTERS

HLT 2010, Riga
07.10.2010.

Digitization projects at NLL

1. ***Heritage-1***
   - Years: 2000.-2006.
   - ~2 million pages
2. ***Periodicals.lv***
   - Years: 2007.-2008.
   - 350 000 pages
   - OCR
3. ***Mass-digitization***
   - Years: 2010.-2012.
   - Periodicals & Books
   - OCR & interactive content

“Heritage – 1”

http://data.lnb.lv/digitala_biblioteka/laikraksti/

“Heritage – 1”

Laikraksts "Jaunā Dienas Lapa", 1905. - 1918. gads

Gadi:

- 1905
- 1906
- 1907
- 1908
- 1909
- 1910
- 1911
- 1912
- 1913
- 1914
- 1915
- 1916
- 1917
- 1918

Apraksts

1914. gads (Nr./datums)

1/2.01.	2/3.01.	3/4.01.	4/7.01.	5/8.01.
6/9.01.	7/10.01.	8/11.01.	9/13.01.	10/14.01.
11/15.01.	12/16.01.	13/17.01.	14/18.01.	15/20.01.
16/21.01.	17/22.01.	18/23.01.	19/24.01.	20/25.01.
21/27.01.	22/28.01.	23/29.01.	24/30.01.	25/31.01.
26/1.02.	27/3.02.	28/4.02.	29/5.02.	30/6.02.
31/7.02.	32/8.02.	33/10.02.	34/11.02.	35/12.02.
36/13.02.	37/14.02.	38/15.02.	39/17.02.	40/18.02.
41/19.02.	42/20.02.	43/21.02.	44/22.02.	45/24.02.
46/25.02.	47/27.02.	48/28.02.	49/1.03.	50/3.03.
51/4.03.	52/5.03.	53/6.03.	54/7.03.	55/8.03.
56/10.03.	57/11.03.	58/12.03.	59/13.03.	60/14.03.
61/15.03.	62/17.03.	63/18.03.	64/19.03.	65/20.03.
66/21.03.	67/22.03.	68/24.03.	69/26.03.	70/27.03.
71/28.03.	72/29.03.	73/31.03.	74/1.04.	75/2.04.
76/3.04.	77/5.04.	78/8.04.	79/9.04.	80/10.04.
81/11.04.	82/12.04.	83/14.04.	84/15.04.	85/16.04.
86/17.04.	87/18.04.	88/19.04.	89/21.04.	90/22.04.
91/23.04.	92/24.04.	93/25.04.	94/26.04.	95/28.04.
96/29.04.	97/30.04.	98/1.05.	99/2.05.	100/3.05.
101/5.05.	102/6.05.	103/7.05.	104/8.05.	105/9.05.
106/10.05.	107/12.05.	108/13.05.	109/14.05.	110/16.05.
111/17.05.	112/19.05.	113/20.05.	114/21.05.	115/22.05.
116/23.05.	117/24.05.	118/27.05.	119/28.05.	120/29.05.
121/30.05.	122/31.05.	123/2.06.	124/3.06.	125/4.06.
126/5.06.	127/6.06.	128/7.06.	129/9.06.	130/10.06.

Periodicals.lv

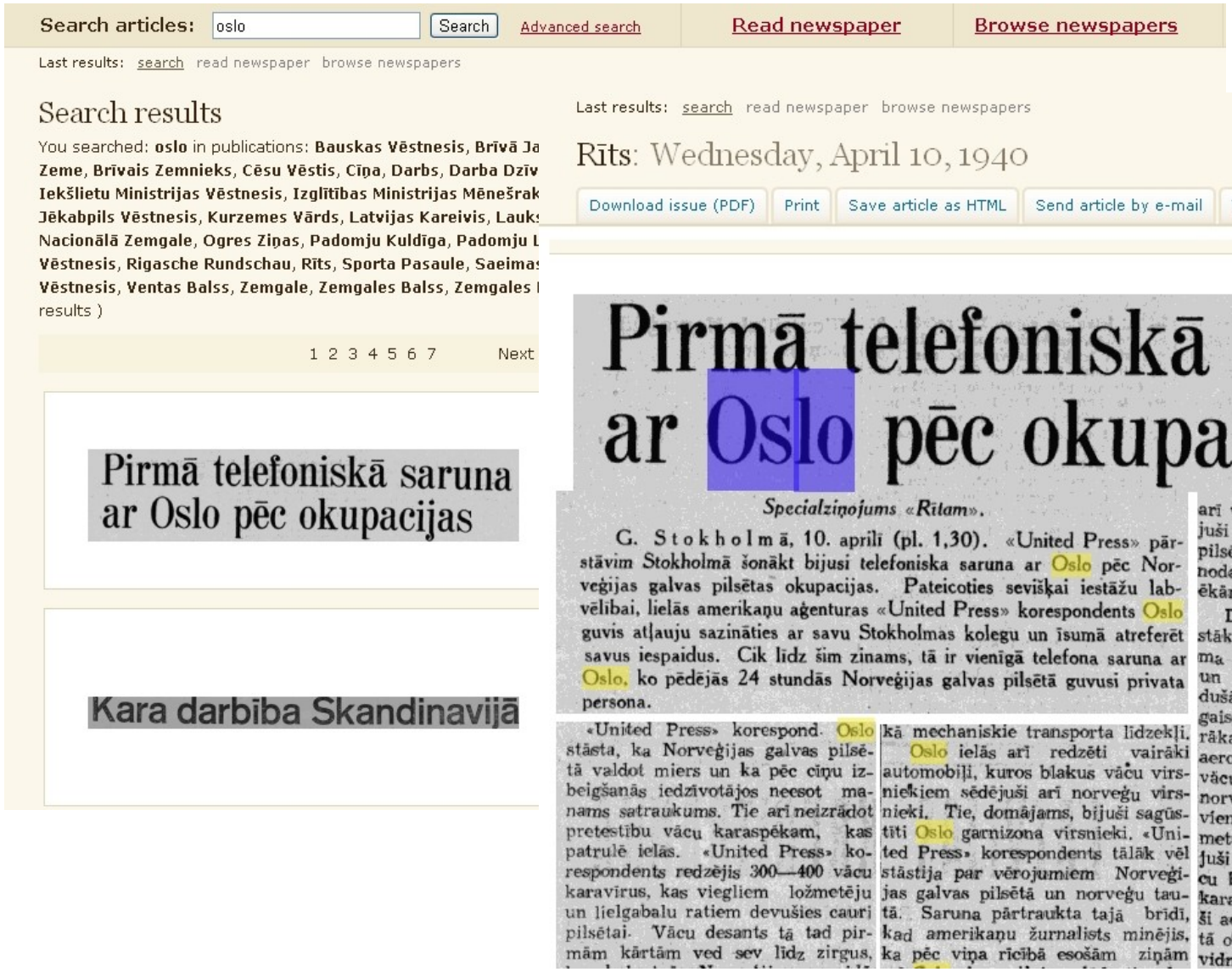
- Time frame: 2007.-2008.
- Scope of the project:
  - 350 000 pages (approx. 45 000 issues)
  - Newspapers issued 1895.-1957.
  - Segmentation & OCR

- Segmentation, OCR, portal by *Olive Software*

Periodicals.lv

http://periodicals.lv

Periodicals.lv



Search articles: oslo Search Advanced search | Read newspaper | Browse newspapers

Last results: search read newspaper browse newspapers

Search results

You searched: **oslo** in publications: **Bauskas Vēstnesis, Brīvā Ja Zeme, Brīvais Zemnieks, Cēsu Vēstis, Cīņa, Darbs, Darba Dzīv Iekšlietu Ministrijas Vēstnesis, Izglītības Ministrijas Mēnešrak Jēkabpils Vēstnesis, Kurzemes Vārds, Latvijas Kareivis, Lauks Nacionālā Zemgale, Ogres Ziņas, Padomju Kuldīga, Padomju L Vēstnesis, Rigasche Rundschau, Rīts, Sporta Pasaule, Saeimas Vēstnesis, Ventas Balss, Zemgale, Zemgales Balss, Zemgales** results)

1 2 3 4 5 6 7 Next

Pirmā telefoniskā saruna ar Oslo pēc okupacijas

Kara darbība Skandinavijā

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)

Pirmā telefoniskā saruna ar Oslo pēc okupacijas

Specializējums «Rītam».

G. Stokholmā, 10. aprilī (pl. 1,30). «United Press» pārstāvim Stokholmā šonākt bijusi telefoniska saruna ar Oslo pēc Norvēģijas galvas pilsētas okupacijas. Pateicoties sevišķai iestāžu labvēlībai, lielās amerikaņu aģenturas «United Press» korespondents Oslo guvis atļauju sazināties ar savu Stokholmas kolegu un īsumā atreferēt savus iespaidus. Cik līdz šim zinams, tā ir vienīgā telefona saruna ar Oslo, ko pēdējās 24 stundās Norvēģijas galvas pilsētā guvusi privata persona.

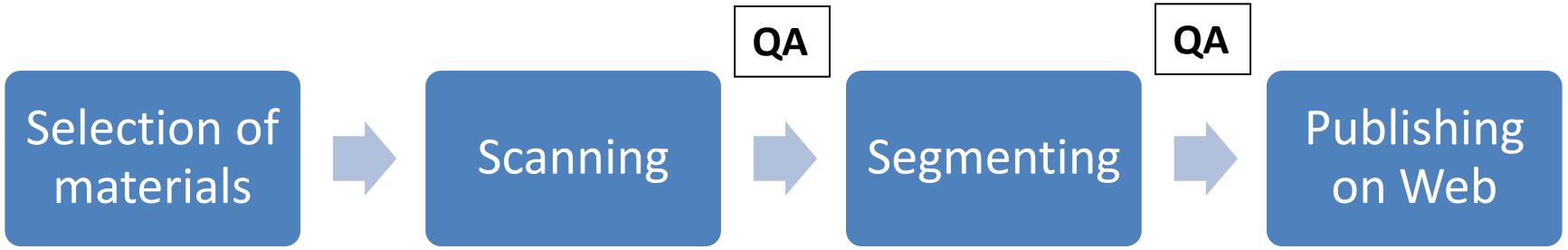
«United Press» korespond. Oslo stāsta, ka Norvēģijas galvas pilsētā valdot miers un ka pēc cīņu izbeigšanās iedzīvotājos neesot manams satraukums. Tie arī neizrādot pretestību vācu karaspēkam, kas patrulē ielās. «United Press» korespondents redzējis 300—400 vācu karavīrus, kas viegliem ložmetēju un lielgabalu ratiem devušies cauri pilsētai. Vācu desants tā tad pirmām kārtām ved sev līdz zirgus, kā mechaniskie transporta līdzekļi. Oslo ielās arī redzēti vairāki automobiļi, kuros blakus vācu virsniekiem sēdējuši arī norveģu virsnieki. Tie, domājams, bijuši sagūstīti Oslo garnizona virsnieki. «United Press» korespondents tālāk vēl stāstīja par vērojumiem Norvēģijas galvas pilsētā un norveģu tautā. Saruna pārtraukta tajā brīdī, kad amerikaņu žurnālists minējis, ka pēc viņa rīcībā esošām ziņām arī vācu karakuģi neesot 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 nolaidušās Oslo apkārtnē un izcēlušas gaisa desantus, kas ieņēmuši vairākas Oslo pilsētas daļas. Oslo aerodromā nolaidušies 30 smagie vācu bumbvedēji, kuriem pretim norveģi varējuši stāties tikai ar vienu pretaviacijas bateriju. Nometuši dažas bumbas un apšaudījuši aerodromu ar ložmetējiem, vācu bumbvedēji izcēluši bruņotus karavīrus, kas īsā cīņā pārvarējuši aerodroma sardzi. Līdzīgā kārtā okupēti arī citi aerodromi Dienvidnorveģijā.

Mass-digitization

- Funded by *European Regional Development Fund*
- Time frame: 2009.-2012.
- Scope of the project:
  - Periodicals: ~2.1 million pages (~700 titles)
  - Books: ~1.4 million pages (~7000 books)
  - Portal of searchable historical texts
  - User interactivity

Digitization process



National Library of Latvia

National Library of Australia, National Library of Netherlands

Scanning

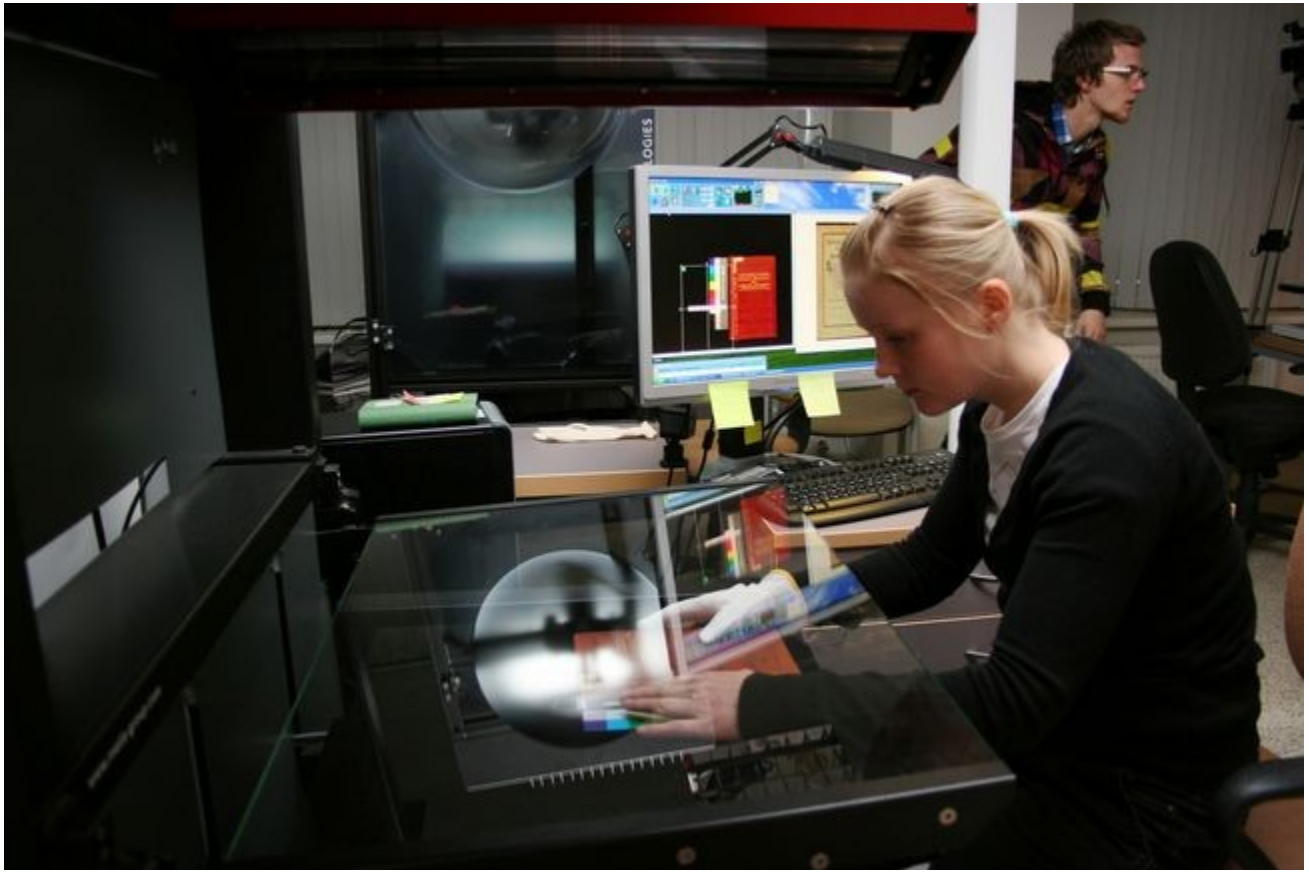
- Done by 

- 1 shipment every 2 weeks:
  - Periodicals: ~46 000 pages
  - Books: ~55 000 pages

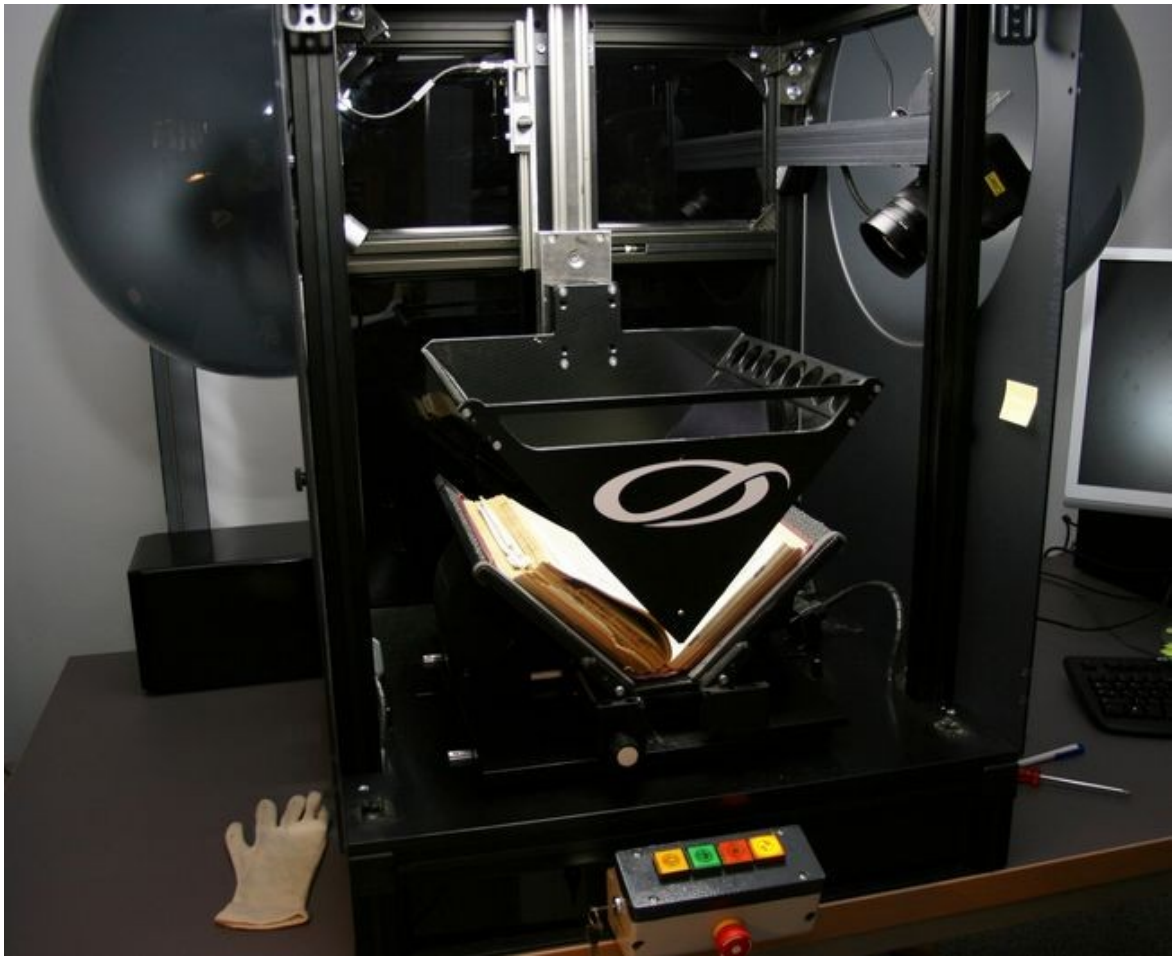
Scanning

Scanning

Scanning



Scanning



Scanning output

- JPEG 2000 file for each page
  - Books, magazines – colour (RGB)
  - Newspapers – Greyscale
  - Resolution: 400 dpi

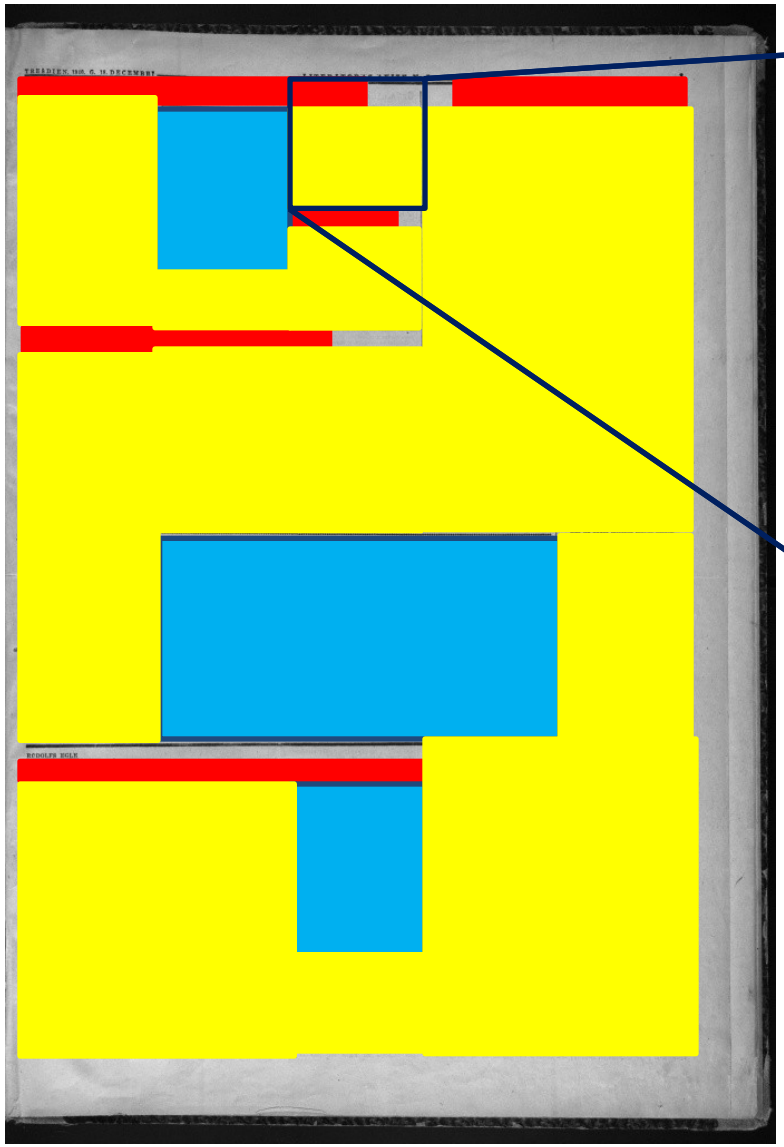
- File size: 3-100 MB

Scanning output

Segmentation

- Identify logical parts of material:
  - Articles/titles
  - Images/captions
  - Authors
  - Tables
  - Advertisments
  - ...
- Text recognition (OCR)

Segmentation



Maksims Goŗkijs Rīgas jūrmalā

1905. gada rudenī ievērojamais krievu proletāriāta rakstnieks Maksims Goŗkijs pavadīja dažas nedēļas Rīgas jūrmalā. Izvairīdamies Pēterburgas žandarmērijas, viņš gribēja te mazliet atpūsties. Bet nepagāja pāris nedēļu, kad Rīgas jūrmalas policijmeistars saņēma no Pēterburgas telegramu nekavējoties arestēt M. Goŗkiju un nosūtīt viņu uz Pēterburgu. Telegramā nebija norādīts pamats; policijmeistars pieprasīja Pēterburgā, kādu Krimināl kodeka pantu piemērot Goŗkijam.

Vidzemes žandarmērijas archīvā, kas atrodas Igaunijas PSR centrālarchīvā Tartū, glabājas plaša, apmēram 40 lappušu bieza Goŗkija lieta.

OCR

Maksims Gorkijs
Rīgas jūrmalā

1905. gada rudenī ievērojamais krievu proletariāta rakstnieks M a k s i m s G o r k i j s pavadīja dažas nedēļas Rīgas jūrmalā...

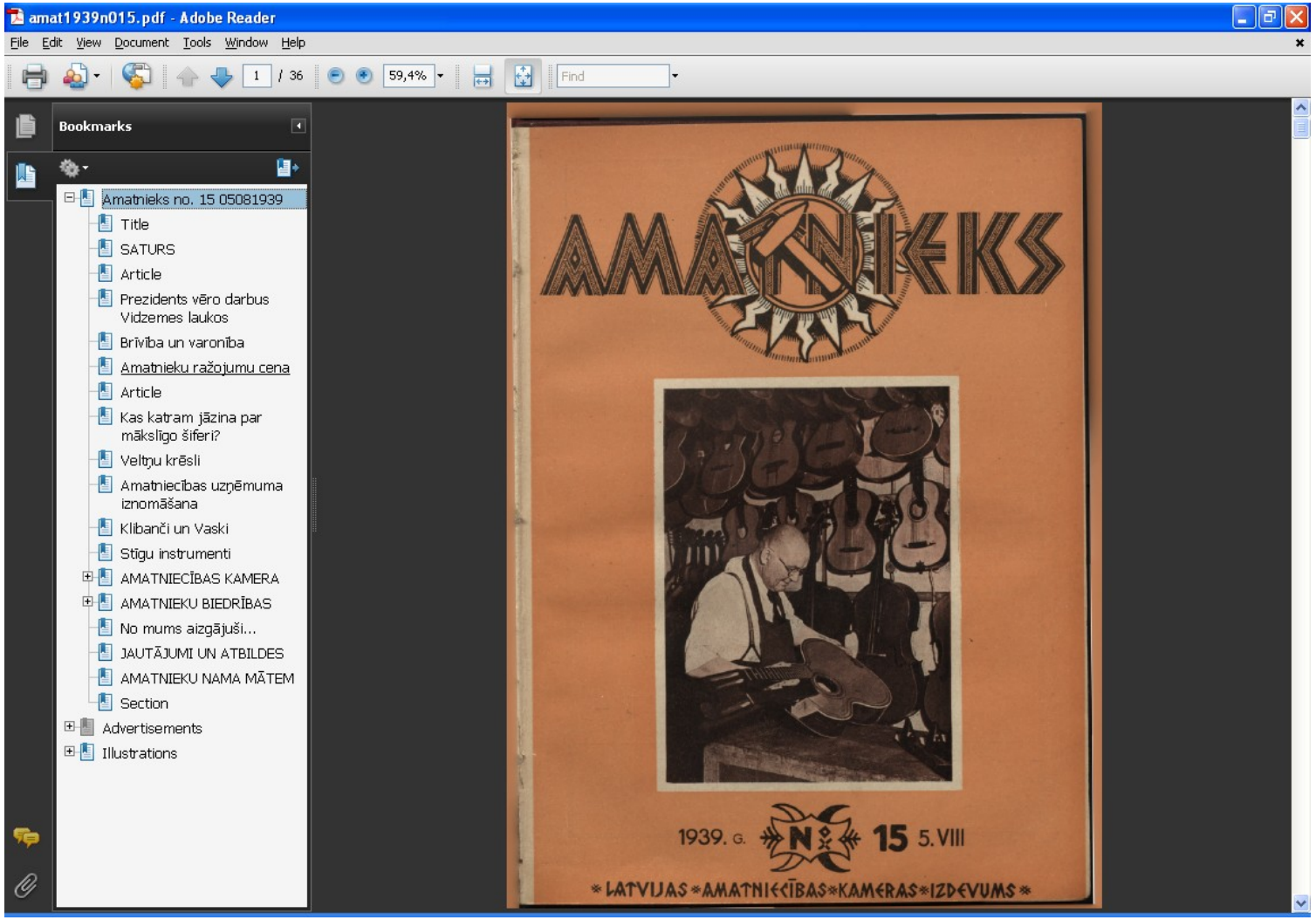
Segmentation output

- Files generated during segmentation:
  - **1 METS** file – for each issue
  - **1 ALTO** file – for each page
  - **1 JPG** file – for each page
  - **1 OCR** file – for each article

Will be used to create the portal

  - **1 PDF** file – for each issue

Segmentation output



PDF with table of contents

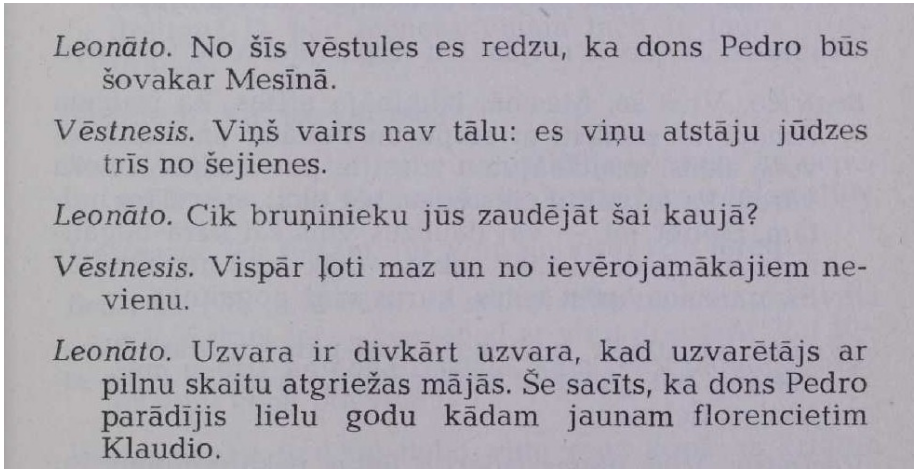
OCR

- Performance of OCR (per character)
  - Modern text – close to 100%
  - Gothic fonts – 80-90%
  - Faded typewriting - <50%

- **Manual correction** of *titles* and *image captions*

OCR performance

- **Modern texts**



Leonāto. No šīs vēstules es redzu, ka dons Pedro būs šovakar Mesīnā.

Vēstnesis. Viņš vairs nav tālu: es viņu atstāju jūdzes trīs no šejienes.

Leonāto. Cik bruņinieku jūs zaudējāt šai kaujā?

Vēstnesis. Vispār ļoti maz un no ievērojamākajiem nevienu.

Leonāto. Uzvara ir divkārt uzvara, kad uzvarētājs ar pilnu skaitu atgriežas mājās. Še sacīts, ka dons Pedro parādījis lielu godu kādam jaunam florencietim Klaudio.

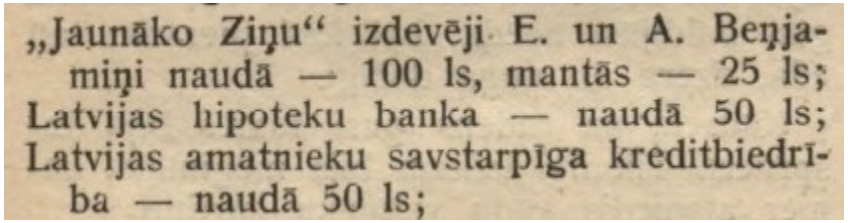
Original

Leonato. No šīs vēstules es redzu, ka dons Pedro bus šovakar Mesīnā.
Vēstnesis. Viņš vairs nav talu: es viņu atstāju jūdzes trīs no šejienes.
Leonato. Cik bruņinieku jus zaudējat šai kauja?
Vēstnesis. Vispār ļoti maz un no ievērojamākajiem nevienu.
Leonāto. Uzvara ir divkārt uzvara, kad uzvarētājs ar pilnu skaitu atgriežas mājās. Še sacīts, ka dons Pedro parādījis lielu godu kādam jaunam florencietim Klaudio.

OCR

Correct characters/total: 396/403 (~98%)

«108% OCR»


„Jaunāko Ziņu“ izdevēji E. un A. Beņjaminņi naudā — 100 ls, mantās — 25 ls; Latvijas hipoteku banka — naudā 50 ls; Latvijas amatnieku savstarpīga kreditbiedrība — naudā 50 ls;

Original

..hipoteku.. ..kreditbiedrība...

Obsolete ortography

„Jaunāko Ziņu" izdevēji E. un A. Beņjaminņi naudā — 100 ls, mantās — 25 ls; Latvijas hipotēku banka — naudā 50 ls; Latvijas amatnieku savstarpīga kredītbiedrība — naudā 50 ls;

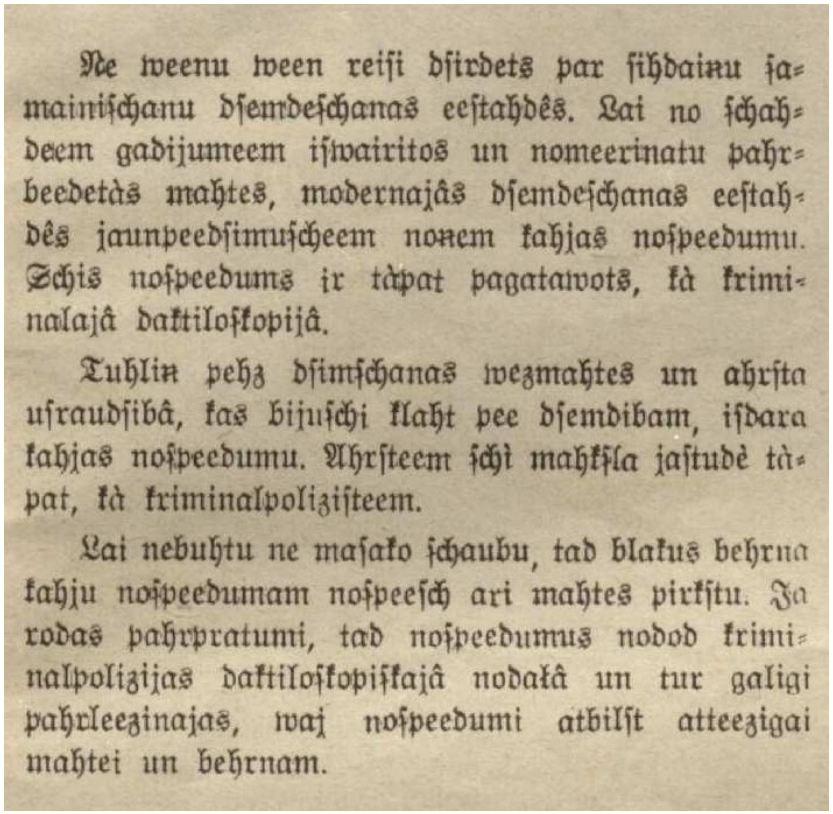
OCR

..hipotēku.. ..kredītbiedrība...

Modern ortography

OCR performance

- **Gothic fonts**



Ne weenu ween reisi dsirdets par sihdainu samainischanu dsemdeschanas eestahdès. Lai no schahdeem gadijumeem iswairitos un nomeerinatu pahrbeedetàs mahtes, modernajâs dsemdeschanas eestahdès jaunpeedsimuscheem nonem kahjas nospeedumu. Schis nospeedums ir tàpat pagatawots, kà kriminalajâ daktiloskopijâ.

Tuhlin pehz dsimschanas wezmahtes un ahrsta usraudsibâ, kas bijuschi klaht pee dsemdibam, isdara kahjas nospeedumu. Ahrsteem schi mahksla jàstudè tàpat, kà kriminalpolizisteem.

Lai nebuhtu ne masako schaubu, tad blakus behrna kahju nospeedumam nospeesch ari mahtes pirkstu. Ja rodas pahrpratumi, tad nospeedumus nodod kriminalpolizijas daktiloskopiskajâ nodalâ un tur galigi pahrleezinajas, waj nospeedumi atbilst atteezigai mahtei un behrnam.

Original

Ne meenu Meen reisi dsirdets par sihdaiņu samainischanu dsemdeschanas eestahdss. Lai no schahdsem gadijumeem iswairitos un nomeerinatu pahrbeedetas mahtes, modernajās dsļnndeschanas eestahdes jaunpeedsimuscheem noņem kahjas nospeedumu. Schis nospcedums ir tāpat pagatamots, ka krimi-- nalaja daktiloskopija.
Tuhliņ pehz dsimschanas mezmahtes un ahrsta usraudsiba, kas bijuschi klaht pee dsemdibam, isdara kahjas nospeedumu. Ahrsteem schi mahksla jāstuḍē' tāpat, kti kriminalpolizisteem.
Lai nebuhtu ne masako schaubu, tad blakus behriui tahju nospcedumam nospeesch ari mahtes pirkstu, Ja rodas pahrpratumi, tad nospeedumus nodod krimi nalpolizijas daktiloskopiskaja nodaļa un tur galipj pahrleezinajas. maj nospeedumi atbilst atteezigai mahtei un behrnam.

OCR

Correct characters/total: 685/739 (~92.7%)

Improvements by NLL

- With ABBYY
  - Improved OCR engine for Latvian gothic fonts
- With Institute of Mathematics and Computer Science
  - Morphologic analysis of historical Latvian texts

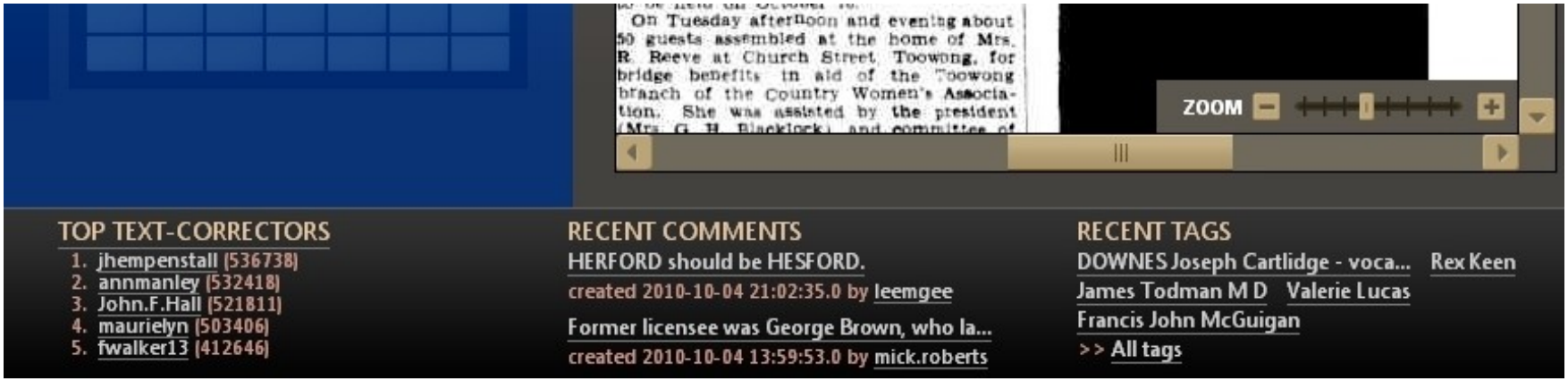
Future perspectives

- National Library of Australia
  - User interaction

http://newspapers.nla.gov.au/ndp/del/home

Future perspectives

- National Library of Australia
  - User interaction



Correcting OCR misspellings **Comments** **Tags**

Future perspectives

- National Library of Australia
  - User interaction

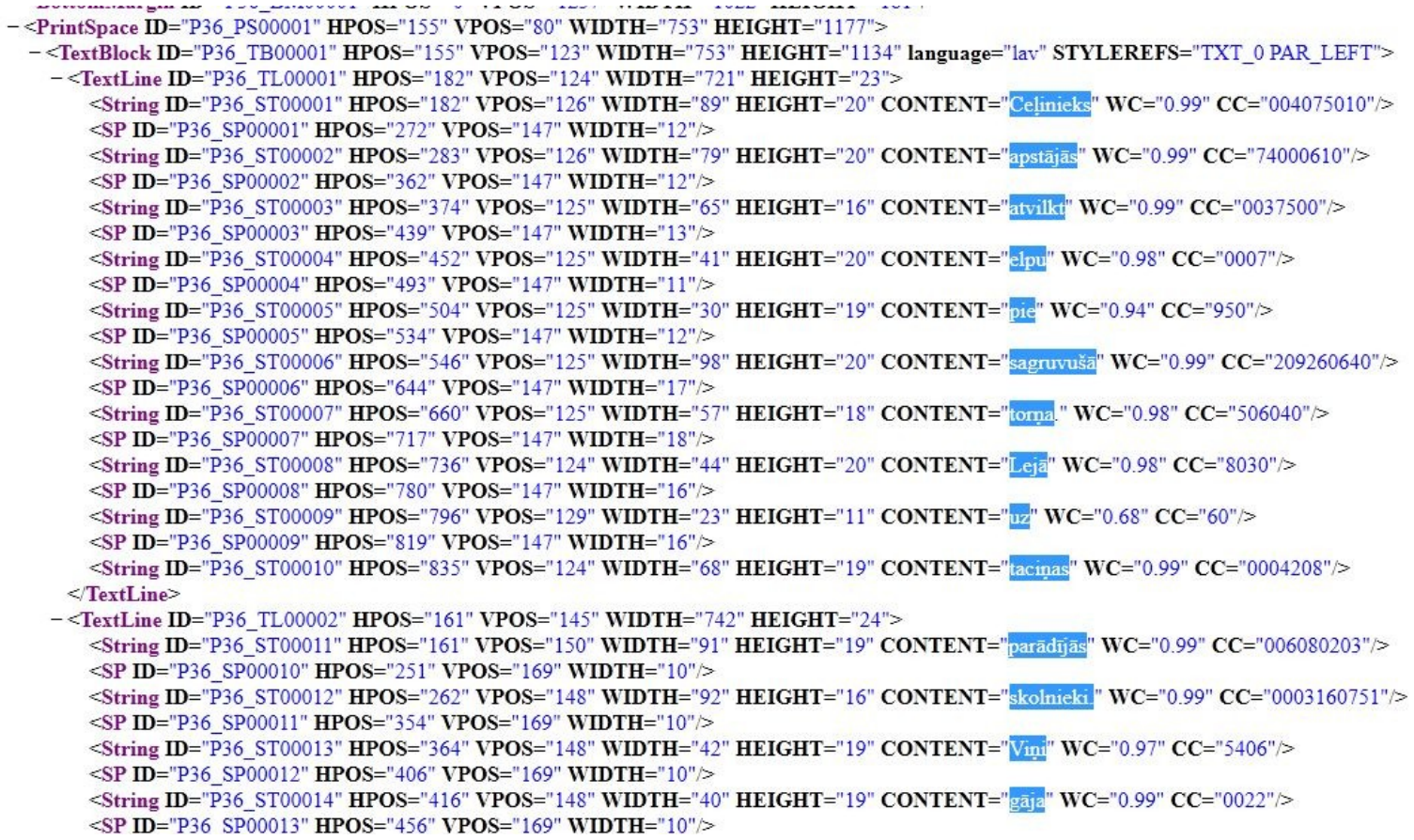
Text corrections

jhempenstall has contributed corrections to **536738** lines; most recently:

Article	Changed	Old Lines	New Lines
THE END OF DEEMING. (FROM THE WEST AUSTRALIAN.)	2010-10-04 11:02:19.0	^ THE END OP DEEMING. \ (FÉOM THB WBBT AUSTRALIAN.) [ON Monday last the curtain fell on the * last act but one in a remarkable career, I trusted the next act will , be the closing	THE END OF DEEMING. (FROM THE WEST AUSTRALIAN.) ON Monday last the curtain fell on the last act but one in a remarkable career, trusted the next act will be the closing

Future perspectives

- Input to Latvian language corpus



```
<PrintSpace ID="P36_PS00001" HPOS="155" VPOS="80" WIDTH="753" HEIGHT="1177">
  <TextBlock ID="P36_TB00001" HPOS="155" VPOS="123" WIDTH="753" HEIGHT="1134" language="lav" STYLEREFS="TXT_0 PAR_LEFT">
    <TextLine ID="P36_TL00001" HPOS="182" VPOS="124" WIDTH="721" HEIGHT="23">
      <String ID="P36_ST00001" HPOS="182" VPOS="126" WIDTH="89" HEIGHT="20" CONTENT="Ceļinieks" WC="0.99" CC="004075010"/>
      <SP ID="P36_SP00001" HPOS="272" VPOS="147" WIDTH="12"/>
      <String ID="P36_ST00002" HPOS="283" VPOS="126" WIDTH="79" HEIGHT="20" CONTENT="apstājās" WC="0.99" CC="74000610"/>
      <SP ID="P36_SP00002" HPOS="362" VPOS="147" WIDTH="12"/>
      <String ID="P36_ST00003" HPOS="374" VPOS="125" WIDTH="65" HEIGHT="16" CONTENT="atvilkt" WC="0.99" CC="0037500"/>
      <SP ID="P36_SP00003" HPOS="439" VPOS="147" WIDTH="13"/>
      <String ID="P36_ST00004" HPOS="452" VPOS="125" WIDTH="41" HEIGHT="20" CONTENT="elpu" WC="0.98" CC="0007"/>
      <SP ID="P36_SP00004" HPOS="493" VPOS="147" WIDTH="11"/>
      <String ID="P36_ST00005" HPOS="504" VPOS="125" WIDTH="30" HEIGHT="19" CONTENT="pie" WC="0.94" CC="950"/>
      <SP ID="P36_SP00005" HPOS="534" VPOS="147" WIDTH="12"/>
      <String ID="P36_ST00006" HPOS="546" VPOS="125" WIDTH="98" HEIGHT="20" CONTENT="sagruvušā" WC="0.99" CC="209260640"/>
      <SP ID="P36_SP00006" HPOS="644" VPOS="147" WIDTH="17"/>
      <String ID="P36_ST00007" HPOS="660" VPOS="125" WIDTH="57" HEIGHT="18" CONTENT="torņa." WC="0.98" CC="506040"/>
      <SP ID="P36_SP00007" HPOS="717" VPOS="147" WIDTH="18"/>
      <String ID="P36_ST00008" HPOS="736" VPOS="124" WIDTH="44" HEIGHT="20" CONTENT="Lejā" WC="0.98" CC="8030"/>
      <SP ID="P36_SP00008" HPOS="780" VPOS="147" WIDTH="16"/>
      <String ID="P36_ST00009" HPOS="796" VPOS="129" WIDTH="23" HEIGHT="11" CONTENT="uz" WC="0.68" CC="60"/>
      <SP ID="P36_SP00009" HPOS="819" VPOS="147" WIDTH="16"/>
      <String ID="P36_ST00010" HPOS="835" VPOS="124" WIDTH="68" HEIGHT="19" CONTENT="taciņas" WC="0.99" CC="0004208"/>
    </TextLine>
    <TextLine ID="P36_TL00002" HPOS="161" VPOS="145" WIDTH="742" HEIGHT="24">
      <String ID="P36_ST00011" HPOS="161" VPOS="150" WIDTH="91" HEIGHT="19" CONTENT="parādījās" WC="0.99" CC="006080203"/>
      <SP ID="P36_SP00010" HPOS="251" VPOS="169" WIDTH="10"/>
      <String ID="P36_ST00012" HPOS="262" VPOS="148" WIDTH="92" HEIGHT="16" CONTENT="skolnieki." WC="0.99" CC="0003160751"/>
      <SP ID="P36_SP00011" HPOS="354" VPOS="169" WIDTH="10"/>
      <String ID="P36_ST00013" HPOS="364" VPOS="148" WIDTH="42" HEIGHT="19" CONTENT="Viņi" WC="0.97" CC="5406"/>
      <SP ID="P36_SP00012" HPOS="406" VPOS="169" WIDTH="10"/>
      <String ID="P36_ST00014" HPOS="416" VPOS="148" WIDTH="40" HEIGHT="19" CONTENT="gāja" WC="0.99" CC="0022"/>
      <SP ID="P36_SP00013" HPOS="456" VPOS="169" WIDTH="10"/>
```

ALTO file

Q & A

arturs.zogla@lnb.lv

jurgis.skilters@lu.lv