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

Arturs Zogla

*Head of Digital Library, National Library of Latvia,*
*arturs.zogla@lnb.lv*

## Abstract

This article gives a general overview of newspaper digitization projects at National Library of Latvia (NLL) and provides details on several experimental solutions that make use of OCRed text of digitized newspapers. Article explains three stages of newspaper digitization initiatives at NLL and how new functionality was added with each stage. It also describes old word "modernization" service and how it was used to process OCR text of newspapers in order to create: a.) Newspaper text corpora; b.) List of Named Entities; c.) Time-sensitive dictionaries. Finally, the article contains suggestions of what a future newspaper portal might look like and in particular describes the *magic glasses* service.

## 1. First attempts at newspaper digitization at the National Library of Latvia

National Library of Latvia has been digitizing newspapers since 2000, when a project titled "Heritage-1" [1] was launched in collaboration with some other libraries in Latvia.

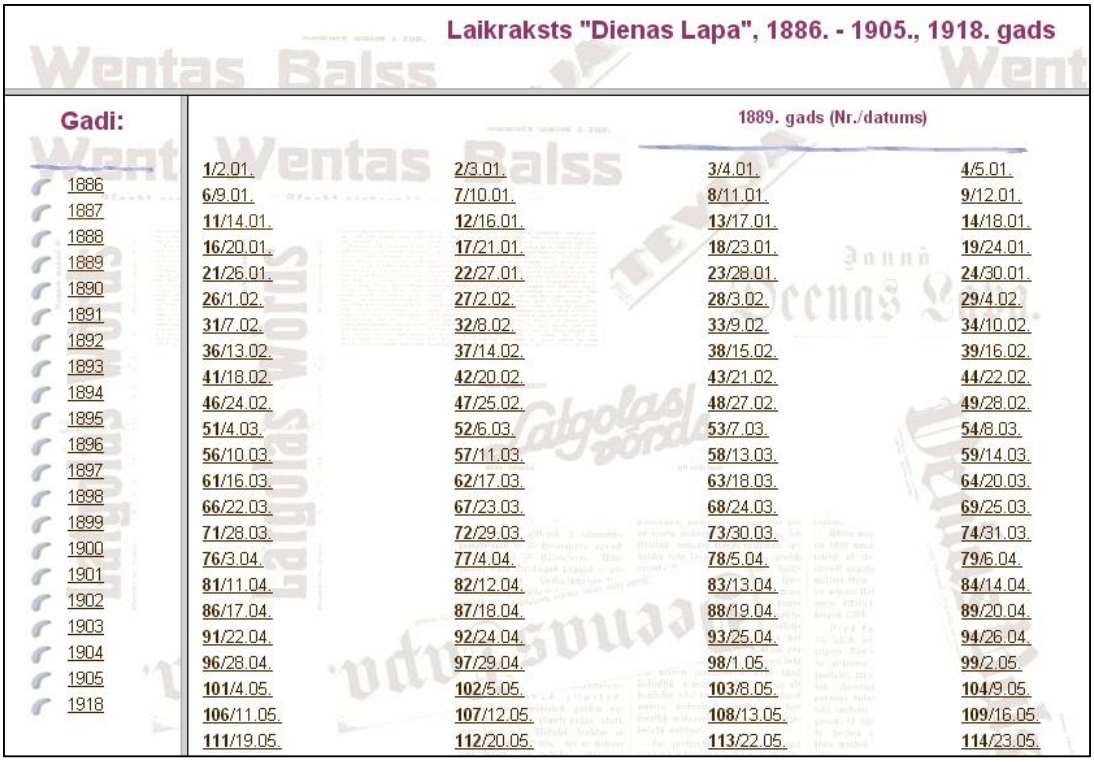



**Image 1 Browsing newspaper issues by date on "Heritage-1" website. Each link in the right frame points to newspaper issue PDF with no OCR.**

Because NLL had no previous experience with digitization, some rules, that today would seem obvious, weren't followed. Image format used in scanning was a simple, low-resolution JPEG instead of a lossless TIFF or JPEG 2000. No OCR was performed on newspapers and only image based PDFs were produced. As a result first newspaper web site had very basic browsing functionality and readers had to know in advance which newspaper issues to look for (Image 1).

Still, even this first project demonstrated two main benefits of newspaper digitization:

- Anyone could access a newspaper archive from their home computer.
- By combining holdings of several libraries, full collections of some newspapers could be created in a digital form that no library on its own held in a physical collection.

In 2006 NLL decided to run OCR on some of the digitized newspapers and create a website with full-text search functionality. NLL investigated an option to digitize already existing microfilms of newspapers. However, first scans revealed poor quality of microfilm images and many samples even contained pages with parts of text cropped. So decision to digitize original newspapers was made. About 350 000 pages were scanned for a pilot project. Post processing of images was outsourced to *Olive Software*, which performed both segmentation/OCR and created a web site with the processed content (Image 2). Segmentation was done on an article level, so with this project NLL went straight from image-only digital newspapers to fully detailed layouts with embedded full text.

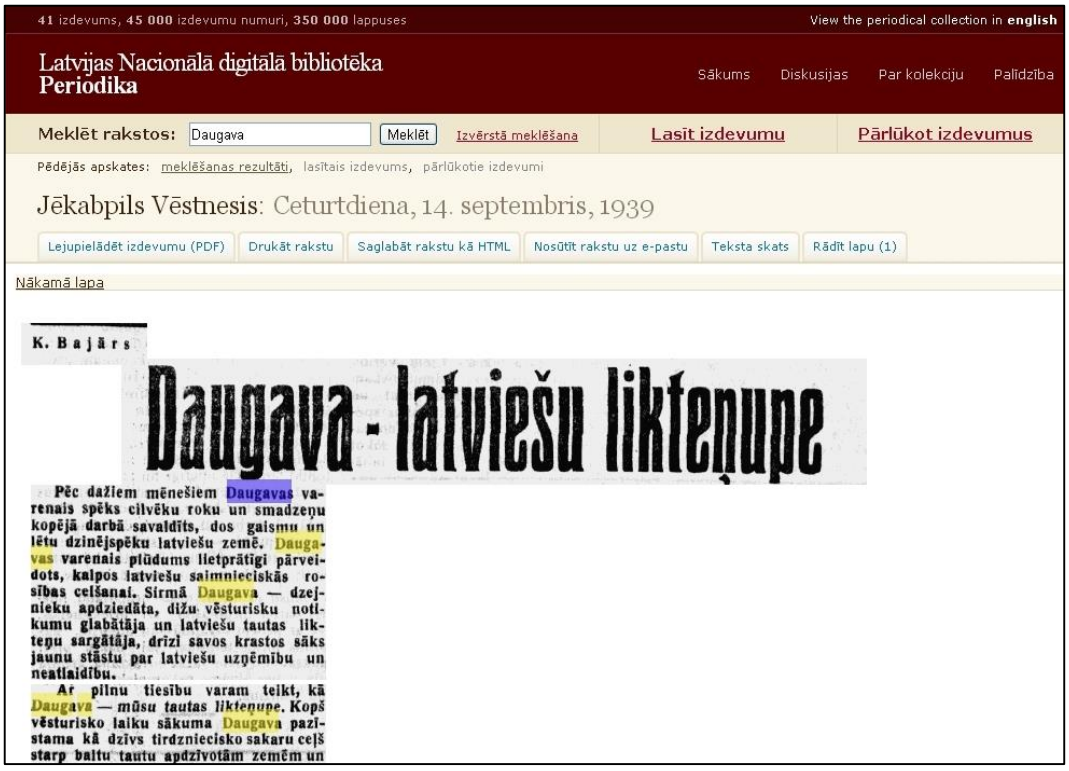



**Image 2 Page in newspaper OCR pilot project website showing an article with searched keyword highlighted**

For the OCR pilot project NLL had to carefully pick newspapers to be processed. On one hand, due to copyright limitations newspapers had to be sufficiently old[1]. On the other hand, because at the time OCR only for antiqua fonts was available, newspapers also had to be recent enough[2]. This limited the potential newspaper collection to periodicals published from about 1920 to 1940.

[1] NLL has adopted a policy to make all newspapers published 70 or more years ago publicly available.
[2] Latvian newspapers used German style gothic fonts until about 1920ies, some newspapers up to 1930ies.

Although contemporary Latvian has characters with diacritics, OCR generally performs very well and for good scans up to 95-98% characters were correctly recognized.

The website created for the OCR pilot project contained following functionality:

- Full text search.
- Advanced search, limiting what elements of newspapers to search, what titles, what period of time, etc.
- Browsing issues by title and date.
- Switching between image and text views.
- Print views of articles.
- Navigating to next and previous issues of newspapers.

The improvements over previous newspaper website were so considerable that page visits to the newspaper website increased up to 20 times. Although the OCR pilot project was a big leap forward from the previous version of newspaper website, it still had some significant limitations both in amount of content and functionality. In 2009 NLL decided to create a periodicals portal that could hold much wider variety of newspapers and would for the first time provide interactive functions.

## 2. Portal of historical newspapers – periodika.lv

In 2009 NLL began a newspaper mass digitization project with an aim to digitize about 2.4 million pages of newspapers over a period of 2 years and to create periodicals portal with advanced search and browse features, as well as interactive functions. From the very beginning it was obvious that both scanning and post-processing of newspapers would have to be outsourced as the expected amount to be processed was estimated at 50 000 pages/week – well beyond NLL's own in-house capabilities.

NLL investigated two different workflows of newspaper digitization. First option was to merge scanning and segmentation into one process. This workflow has already been used by Koninklijke Bibliotheek in Netherlands [2] and National Library of Australia [3]. Second option was to keep scanning and segmentation processes separated. NLL decided to separate scanning and segmentation and these tasks were outsourced to two different companies. This allowed more control over the process, because deliverables were received in smaller, more manageable packages. Also segmentation partner provided additional quality control of the scanning process as each page missing or damaged during scanning was reported to NLL.

Scanning was performed at a local company *Lattelecom BPO* with following parameters:

- Format: JPEG 2000, Greyscale, 400 dpi
- Allowed rotation of text: less than 1 degree
- Average file size of A2 page: 30-100 MB

Segmentation was performed by a local company *LETA* in collaboration with *CCS* and produced METS/ALTO files. Following tasks had to be performed:

- Identification of layout elements: main text, titles, subtitles, images, tables, captions, advertisements[3].
- Manual correction of OCR errors: required for titles and image captions only. In case of very long titles or captions it was acceptable to correct only Named Entities.

[3] On advertisment pages it wasn't required to identify all individual advertisments. It was sufficient to mark all page as an advertisment zone.

- Import of descriptive metadata prepared by NLL for each newspaper issue.

Quality tests on segmented content revealed that OCR for antiqua typeface produces excellent results and many blocks of text were recognized with 100% accuracy. There was even one curious case where a paragraph of text was OCRed with 108% accuracy as OCR engines had "corrected" some typos in the original text[4]. OCR results for gothic typeface were, however, considerably lower – about 90% of characters on average were correctly recognized.

Because no existing solution provided all the functionality NLL had envisioned, portal of periodicals was developed by a local IT company *Datakom* mostly from scratch, but a component for viewing individual newspaper issues created by National Library of Luxembourg was used and adapted for the specific METS/ALTO profile NLL used [4].

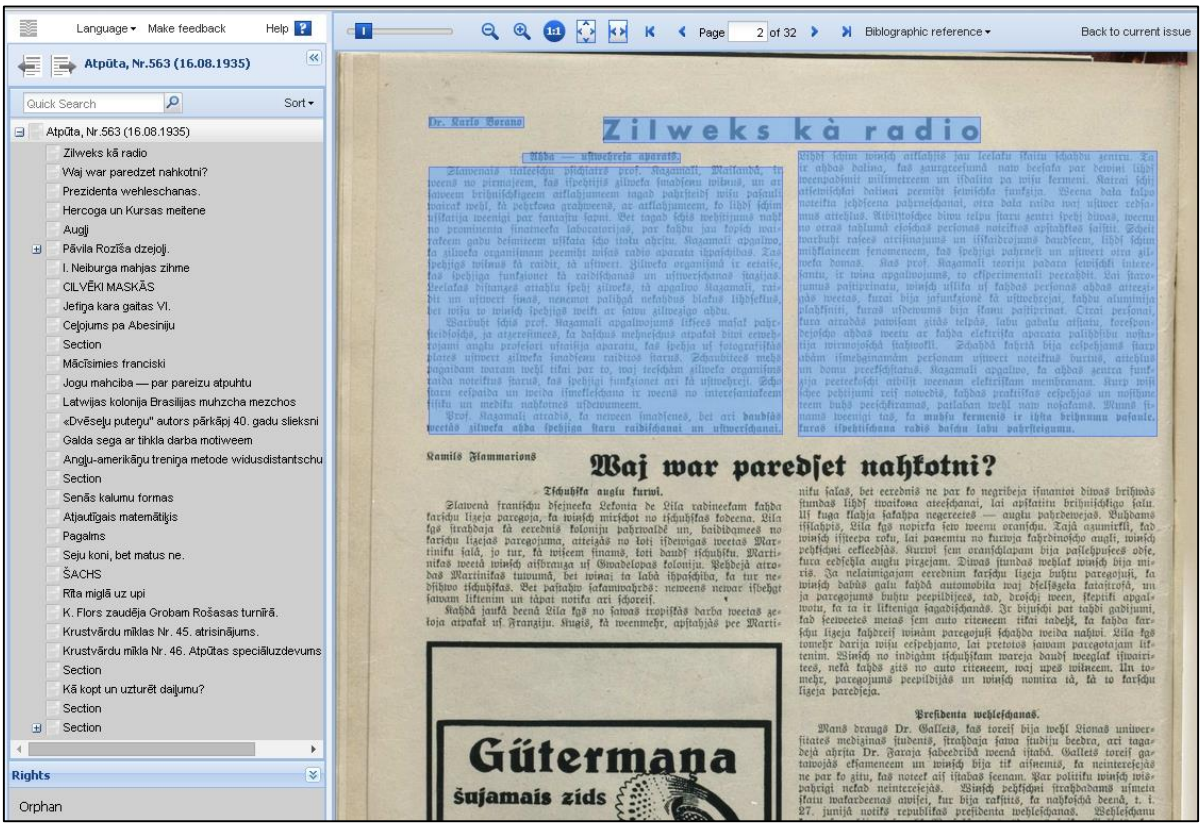



**Image 3 Periodicals viewer developed by National Library of Luxembourg and adapted by National Library of Latvia**

This was also first time NLL included interactive functions in a periodical's portal. Following interactive features were developed:

- Adding comments to articles
- Favouriting articles
- Correcting OCR text
- Sharing content on social networks

NLL had high hopes for OCR correction feature, but so far there has been no considerable amounts corrected. By June 2014, the most active user had corrected 5 678 lines of text, while the next most active – just 636 lines.

[4] In reality all corrected typos from the original text would be considered mistakes as it is NLL's policy to keep all existing errors in the OCR text as well.

Two most likely reasons for low activity in OCR correction are:

1. *Relatively high quality of OCR.* Most antiqua texts are OCRed with almost 100% precision. OCR levels are much lower for gothic typeface, but not many modern readers can read these texts and there is still a discussion, how should texts with deprecated orthography be corrected.
2. *Lack of publicity of OCR correction features.* So far NLL hasn't advertised the OCR correction and those users that have used this feature, have discovered it on their own. A public initiative with a targeted content for corrections would definitely produce higher amounts of corrected text.

There are, however, interesting cases where digitized newspapers have been used by third parties and new websites have been created based on this material. One such website was "Baricadopedia" [5] which collects materials on Latvia's struggle for independence in late 1980ies. NLL provided newspapers from this period and developers of the website manually corrected OCR text of each article and added several tags as well.

«Palikšu partijā»

Arvīds Dauders, Viestarts Gailītis

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

ziedot!

**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 iedzīvotajiem, 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 paliksit?**

— Es palikšu partijā.

**— Iemesli?**

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

«Palikšu partijā»

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

**Image 4 Screenshot of an article on "Barikadopedia" website showing a manually corrected text, preview image of the original article and tags**

Another unexpected example of how newspaper portal is used is to resolve disputes on some linguistic issues. Even professional linguists sometimes post on their Twitter accounts tweets about how some phrase or words have been used in past and prove to others that a particular phrase is older than some might think, because old enough articles can be found containing this phrase.

## 3. Old word modernization service

A periodicals portal could be considered the final stage in newspaper digitization workflow. The next steps and doing actual research is up to readers. However, because OCRed text of newspapers

provides a huge digital text corpus, NLL decided to take few further steps and investigate additional analysis of digitized texts.

By doing QA of digitized texts NLL concluded that OCR quality varies in different newspapers. It was particularly bad for Gothic script. On average Gothic typeface script was OCRed with quality levels of 90% characters recognized correctly, which sometimes meant less than 50% words without OCR errors. But even those words that were recognized correctly were written in an old-style orthography that modern readers often are not familiar with. Yet worse, some obsolete words were last consistently used 100 or more years ago and are completely unrecognizable to most readers.

It was necessary to develop a service that would take as an input an OCRed word in old orthography with possible OCR errors and would:

- automatically correct OCR errors;
- transliterate the word into contemporary orthography;
- in case of obsolete words, would look up a contemporary equivalent;
- look up an explanation of the word from built-in or external dictionaries.

Such a tool was developed in collaboration with Institute of Mathematics and Computer Science and it performed two tasks:

1. ***Transliteration.*** This included both correcting OCR errors and "updating" orthography.
2. ***Normalization.*** Because of rich morphology of Latvian language words in texts are often not in their normal form. Before looking for a word in dictionary, it had to be normalized to single, nominative [6].

Transliteration was rule based where rules could be mandatory, optional or in some cases context-sensitive. Some of the most typical rules were:

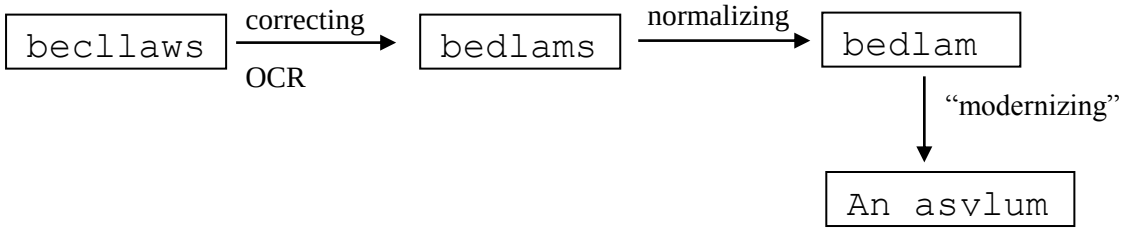
| Rule | Type | Comment |
|---|---|---|
| **m → w** | OCR error / optional | In Gothic script letters "m" and "w" are so similar that even humans without context might confuse these letters. |
| **f → s** | OCR error / optional | Letter "f" is so rare in Latvian, that almost every time it appears in a word, it is in fact incorrectly OCRed letter "s" which in Latvian Gothic is very similar to "f". |
| **w → v** | Orthography / mandatory | Contemporary Latvian alphabet doesn't have letter "w". It was substituted with letter "v" at the beginning of the 20th century. |
| **ah → ā** | Orthography / mandatory | In Latvian Gothic script sound [a:] (like in *"smart"*) was represented with two letters "ah". With introduction of diacritics, this changed to a single letter "ā" in modern Latvian. Extremely rarely there might be a letter combination "ah" even in modern Latvian, but in 99% cases this rule can and will be applied. |
| **tsch → č** | Orthography / mandatory | In some cases even up to 4 letters in Latvian Gothic script became just one letter in modern Latvian. Letter combination "tsch" is impossible in modern Latvian so this rule can be applied in 100% cases. |
| **ee → ē**<br>**ee → ie**<br>**ee → ee** | Orthography / context-sensitive | Some letter combinations can only be transliterated looking at surrounding letters and wider context. |

**Table 1 Typical rules of old word "modernization" service**

Because some of the rules are optional one has to consider both applying and ignoring the rule. This obviously leads to many word variants by the end of rule application chain. The more optional rules can be applied for a particular word, the more variants are generated in the end.

When all rules are applied and final words are normalized, tool tries to look it up in several built-in dictionaries. The output of this tool is a list of corrected and "modernized" words with trust values assigned to each word variant. If a word is generated with application of only mandatory rules and the resulting word can be found in a dictionary, it will get a high score. On the other hand, if some optional rules are applied and the result can't be found in a dictionary, this word variant will get a very low rating.

A simple example of how this would work for an English word is given below:




Detailed analysis of results has concluded that on average tool generates 2.89 word variants for each initial input word with a 92.45% probability that the correct result is among these variants. In some sense this tool provides an advanced spell checking that involves not only finding errors in words, but also word "modernization". And like in traditional spell checking, several word variants provide a list for end-user to choose from.

This tool would be very useful on its own, however NLL decided to use it to create even more sophisticated solutions.

## 4. Experimental computer linguistics solutions at NLL

NLL intends to open up its data and make it freely accessible on web although it's not always possible to give full access to all data. For example, due to copyright restrictions libraries can only freely publish public domain texts. However, it's possible to extract some information even from the protected texts and present it as open data.

NLL decided to perform several computer linguistics experiments on digitized newspapers in order to demonstrate a potential of using data mining techniques on its digital collections. Because of OCR errors and rich variety of documents, at least some pre-processing of text was required. Word "modernization" tool described in previous chapter was particularly useful as most faulty words could be fixed with this tool.

First solution was a **newspaper text corpus**, which essentially allows users to run full text search queries. Much like in newspaper portal itself, but a corpus query building language could be used letting users construct much more complex queries. For example, users can specify proximity of keywords in a phrase. The text corpus contains 4.5 billion tokens and so far is the biggest publicly available text corpus for Latvian.

Second solution was automated **Named Entity Recognition** within newspaper text. Initially this seemed to be an easy task for Latvian texts as only named entities begin with a capital letter in

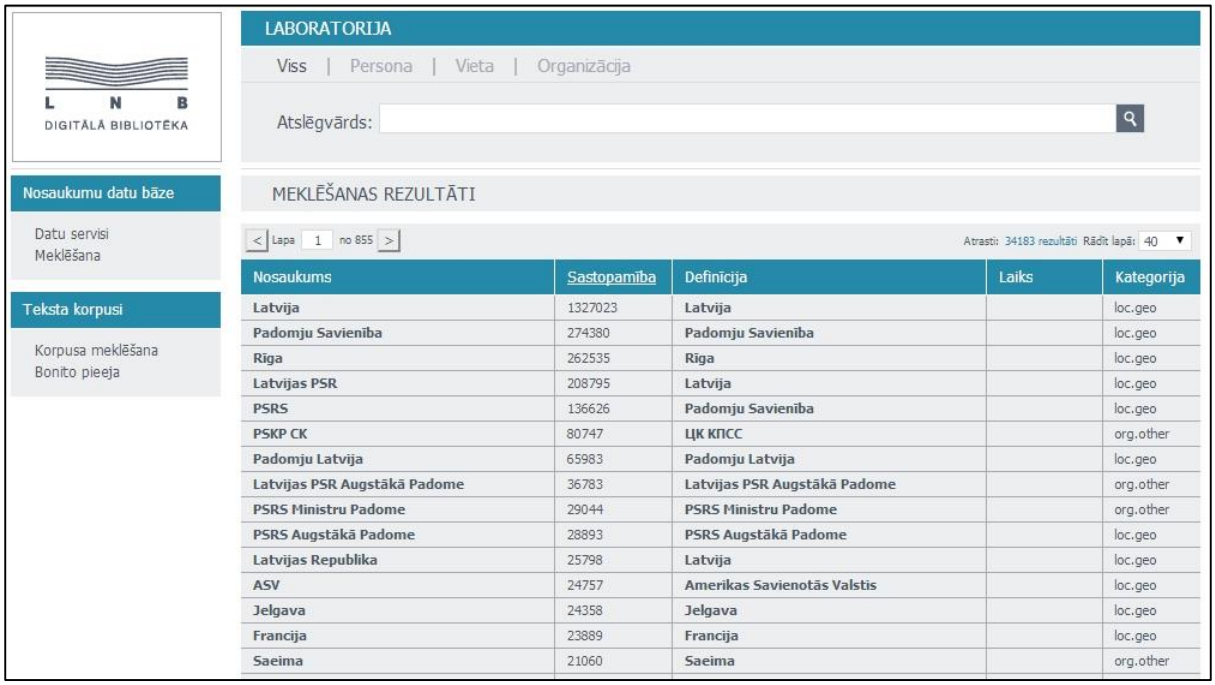
Latvian (except for the first word in a sentence), but most challenges arose from multiple word named entities. For example, in a sentence

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

one obvious named entity would be *"Minister of Culture"*. However, it is non-trivial even for humans to decide how to tag named entities at the beginning of the sentence. Is it the entire phrase *"President of Republic of Latvia Guntis Ulmanis"* or just *"Guntis Ulmanis"*? Should *"Republic of Latvia"* be tagged as an independent entity? Even for the named entity *"Minister of Culture"* a decision must be made whether an implicitly implied *"of Republic of Latvia"* should be added to the named entity.

NLL together with Institute of Mathematics and Computer Science (IMCS) tagged a training set of about 150 000 words from different time periods and this provided ground truth samples for NER tools. The chosen taxonomy consisted of 7 types: person, location, organization, facility, event, product and time. Types were further divided into 21 subtypes [7].

Another input for NER tool were several thesauri created by NLL. Thesauri of persons, institutions and places were used to link several representations of the same entity. So, for example named entities "UN" and "United Nations" would be considered a single entity with two representations.




| Nosaukums | Sastopamība | Definicija | 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 | LĶK 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 |

**Image 5 List of named entities recognized in Latvian newspapers sorted by number of times they appear**

NER tool was based on Stanford CFR classifier. It was ran on entire text corpus of digitized newspapers and only those named entities that were identified at least 10 times were collected. As a result over 26 000 named entities were identified (Image 5) in a text corpus consisting of 4.5 billion tokens. Further studies should reveal amount of both false positives and false negatives within this dataset. Initial manual evaluation of results already revealed that most mistakes are made identifying multiple word organization names. Still, for high quality OCR material when counting as correct only those entities that were identified precisely with the same type and same boundaries as identified by human operators, precision of over 80% was observed.

Finally, NLL together with IMCS created **a time-sensitive dictionary** web service, which would provide assistance in interpreting historic texts. Main tasks of a time-sensitive dictionary include providing:

1. Additional information on implicit named entities, based on article's time stamp. For example, the phrase "president of USA" in a year 1959 article would be deciphered as *John F. Kennedy*, but in a year 2014 article – as *Barack Obama*.
2. Modern version of a known named entity. For example, word "Bombay" in a 1970ies article would be translated into modern version "Mumbai";
3. Interpreting measurement units that have changed over time. This could be particularly useful for some measurement units from imperial unit system.

The idea to create time-sensitive dictionaries actually originated from frequent renaming of streets in Riga[5]. In 20th century alone almost all streets have been renamed at least 3 times. Because newspapers are rich in addresses, especially on advertisement pages, often readers will encounter old street names with no idea what part of modern day Riga these places are located in.

NLL used the fact that it had access to time-sensitive thesaurus of street names in Riga and integrated this into time-sensitive dictionary web service, which would take as input a street name, a year it was mentioned in and would return the current name of the street.

As an extreme example for use of time-sensitive dictionaries we can imagine an integration of daily exchange rates for some particular currency. This would provide different interpretations of "a pound" that would be mentioned in a year 1984 and in a year 2014 articles.

Results of all these computer linguistics experiments are published on-line at NLL's Labs site [8][6].

## 5. Functionality of a futuristic periodicals portal

All of the linguistic solutions described in previous chapters although experimental for now might be integrated into periodicals portal, providing an enhanced user experience. We can imagine a futuristic periodicals portal that besides simply presenting digitized newspapers would also provide a so called *magic glasses* service. Without *magic glasses* on users would only see OCR text with all its flaws: OCR errors, old and unrecognizable words, unidentified persons, places, etc.

With *magic glasses* on following changes might be performed on OCR text:

- Correction of OCR errors.
- "Modernization" of language, i.e., correcting obsolete orthography.
- Explanation of named entities. Both implicit and explicit ones. Explanations might even be retrieved from external sources, like *Wikipedia*.
- Explanation of idioms.
- Substitution of old addresses to modern equivalents.
- Translation of the text to native tongue of user.

In the optimistic scenario such a service might take a medieval text written in fracture script in Latin, understandable only to narrow group of scholars and translate it into modern day English with every obsolete concept, every person name, event and place name explained with links to corresponding *Wikipedia* articles.

A simple example of this would be a following sentence in Old English with OCR errors:

[5] Capital of Latvia
[6] At the time of writing this article, site has a Latvian-only interface.

"On Cyres cyninges dagum wregdon þa Babilomscan þone witegan Daniel, for þæm þe he hiera deofolgield towearp."

translated into:

"In the days of King Cyrus <Cyrus the Great, founder of the Achaemenid Empire. http://en.wikipedia.org/wiki/Cyrus_the_Great>, the Babylonians <*Babylon was originally a Semitic Akkadian city dating from the period of the Akkadian Empire c. 2300 BC*> accused the prophet Daniel <*Daniel. was one of several children taken into Babylonian captivity where they were educated in Chaldean thought*>, because he overthrew their idol."

There are three main scenarios of how to implement a *magic glasses* service:

1. Pre-process all text with the *magic glasses* service once and later, when activated, it would simply retrieve the enhanced version of particular piece of text
2. Crowd-source adding all enhancements, but this even in the best case scenario, would provide improvements only to some parts of text archive while other parts would remain unchanged.
3. Run the *magic glasses* service on-the-fly.

The final option is probably favourable. It might take longer to load the enhanced text, but it would always use latest updates of the service and could also use the most recent information from external sources.

There already exist few notable examples of digitization projects that have implemented at least a simple version of *magic glasses* service [9], [10], but further improvements are certainly possible.

## References


[1] "Heritage-1" – periodicals website. Available on-line at: http://data.lnb.lv/digitala_biblioteka/laikraksti/

[2] Edwin Klijn. The quality of quantity: Newspaper digitization at the Koninklijke Bibliotheek. IFLA (2009). Available on-line at: http://conference.ifla.org/past-wlic/2009/99-klijn-en.pdf

[3] Australian Newspaper Digitisation Program. Available on-line at: http://www.nla.gov.au/content/newspaper-digitisation-program

[4] periodika.lv – periodicals portal. Available on-line at: http://periodika.lv/

[5] Barikadopēdija – digital collection. Available on-line at: http://www.barikadopedija.lv/

[6] Lauma Pretkalniņa, Pēteris Paikens, Normunds Grūzītis, et.al. Making Historical Latvian Texts More Intelligible to Contemporary Readers. *Workshop on Adaptation of Language Resources and Tools for Processing Cultural Heritage Objects* (2012).

[7] Peteris Paikens, Ilze Auzina, Ginta Garkaje and Madara Paegle. Towards named entity annotation of Latvian National Library corpus. *Human Language Technologies – The Baltic Perspective* (2012), p. 169-175.

[8] Laboratorija – collection of computer linguistics tools by National Library of Latvia. Available on-line at: http://laboratorija.lndb.lv/

[9] The Newton project. Available on-line at: http://www.newtonproject.sussex.ac.uk/

[10] The Great Parchment book project. Available on-line at: http://www.greatparchmentbook.org/