An incremental approach to retrodigitization

Janusz S. Bień, Joanna Bilińska, Mateusz Sarnecki
(presented by Mateusz Sarnecki)

University of Warsaw

19 July 2014, Bolzano
European Network of e-Lexicography WG2 meeting
Retro-digitized dictionaries

Dictionaries as corpora

Available at korpusy.klf.uw.edu.pl

- Dictionary of 16-th Century Polish, under development (1966–2025?), currently 36 volumes, c. 20,000 pages
- Linde's Dictionary, 2nd edition (1854–1861), 6 volumes, c. 5,000 pages
- "Warsaw Dictionary" (1900–1927), 8 volumes, c. 8,000 pages
- Geographical Dictionary of the Kingdom of Poland and Other Slavic Countries (1880–1902), 15 volumes, c. 15,000 pages

He gives twice who gives quickly

Improving the quality without changing the software on the server side
A prerequisite: a graphical representation

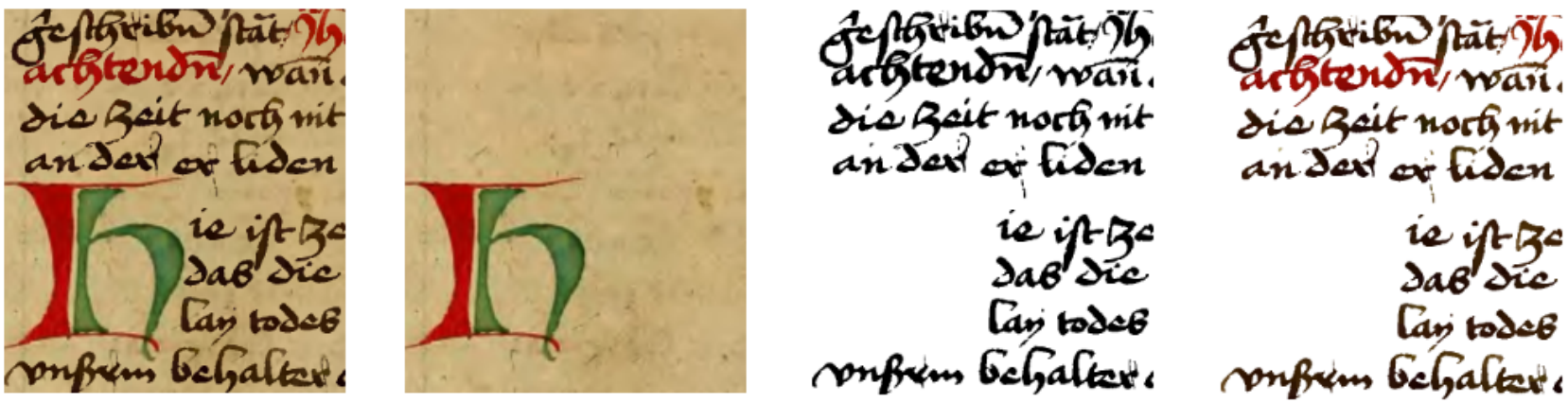
Several independent dimensions of improvements

- Graphical annotations
- Plain text representation
- Tagged text representation
- Indexes

Serving original scans

DjVu technology

DjVu is an image compression technique, a document format, and a software platform for delivering documents images over the Internet (Le Cun et al. 2001)



https://ia600503.us.archive.org/19/items/dasleidenunsersh00bras/dasleidenunsersh00bras.djvu

DjVu

Some advantages

Fast access to a random page without additional server software
Fast zooming from the default resolution to the maximal one
Progressive loading of the file
Full support for URL options
Small file sizes owing to layer separation and compression

Some disadvantages

Need for additional user software, e.g. a browser plugin

Annotating scans

- Always visible, expandable:
  - places affected by original errata
  - places affected by editorial corrections
- “Tooltips”
  - explanation of abbreviations, named entities
    Per: Pressburg, Prezburg, Przezburg — Bratysława (Bratislava)
- Highlighting text areas with colour
  - search hits! root tables?
- Hyperlinks
  - References...
  - Links to digitized sources
  - Links to other dictionaries
- Outlines

From dirty OCR to Ground-Truth text

More than one level of textual representation

Intentional and incidental character substitutions

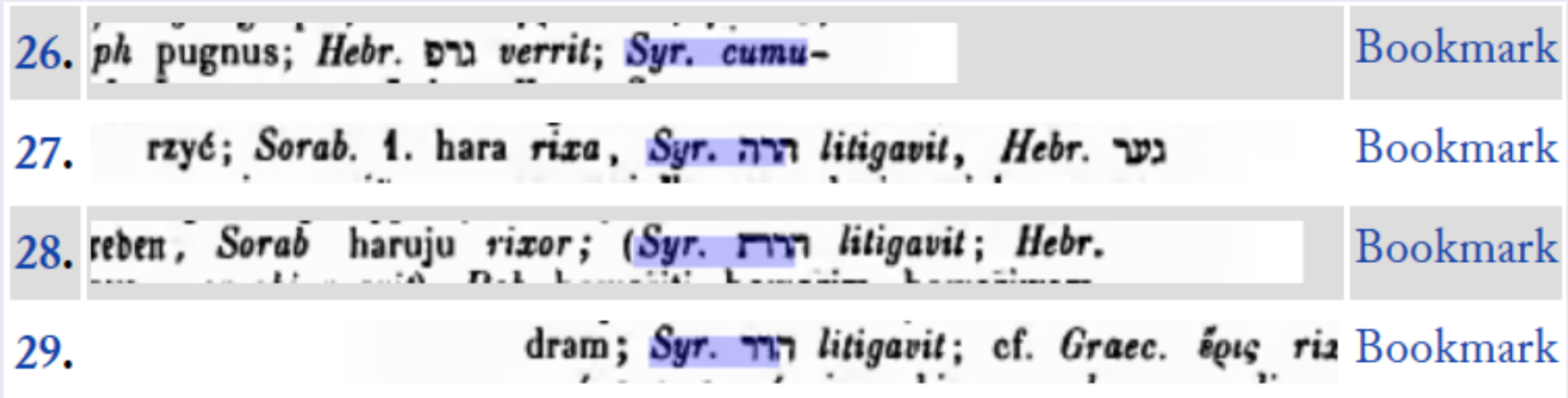
Melabija, *zam.* podanego tam objaśnienia *połóż:* [z *greek.* μελάμβιος = nigram i. e. obscuram ducens vitam. *Hesych.* — 3].

ABREWIACYA, yi. f. *Bosn.* kratcina, *Ross.* сокращенность, словотитла, *Eccl.* скоропись, [ськращєніє, z.] skrócenie, pismo prędkie z tytlami, skoropis, die Abkürzung, Verkürzung, Abbreviatur. Znajdują się w pismach polskich skrócenia czyli abrewiacye, tak przez opuszczenie głosek, jako przez zamianę ich na figury. *Kopcz. Gr.* 1. 29.

Text tagging

Poliqarp for DjVu

Query: `Syr "\." "[^([:digit:])(Ziel)].*"`



From trivial to sophisticated markup

Poliqarp 2

Multiple transcriptions, hyphenation reconstructed

Text tagging

Markup

Possible tagset:

```
[attr]
lang = und pl de ru ...
script = latn latf cyrl ...
series = medium bold
shape = upright italic
wconf = 0 1 2 3 4 5 6 7 8 9
[pos]
ign = lang script series shape wconf
```

Problematic language abbreviations, e.g. Linde's *Carn* for *Carniolice* 'Carniolan' (a dialect group of Slovenian)


1. wotpadnencz; *Carn.* odstopnék; *Vind. et Bosn.* odmetnik; *Croat.* odmetnik, odverzitel; *Rag.* odmetník, raz-

Examples of indexes

Indexes based on running heads

Thesaurus polono-latino-graecus, 1621
Polish-German-French dictionary, 1764

Entry index

reverse index for Linde's dictionary

Auxiliary indexes in Linde

Geographical designations
Foreign words (Italian words in progress)

Relevant publications 1/2

Bień, Janusz S. (2009) *Digitalizing dictionaries of Polish*. In: Methods of Lexical Analysis: Theoretical assumption and practical applications. Wydawnictwo Uniwersytetu w Białymstoku, Białystok, pp. 37-45.
http://bc.klf.uw.edu.pl/71/

Bień, Janusz S. (in press) *The IMPACT project Polish Ground-Truth texts as a DjVu corpus*. Cognitive Studies | Études Cognitives.
http://bc.klf.uw.edu.pl/354/

Bilińska, Joanna A. (2011) *Describing Linde's Dictionary of Polish for digitalisation purposes*. In: Electronic lexicography in the 21st century: new applications for new users (eLex 2011). Trojina, Institute for Applied Slovene Studies, pp. 43-51.
http://bc.klf.uw.edu.pl/240/

Bilińska, Joanna A. (2013) Analiza i leksykograficzny opis struktury słownika Lindego na potrzeby dygitalizacji. [Analysis and lexicographic description of Linde's dictionary for digitization purposes.] PhD thesis, University of Warsaw.
http://bc.klf.uw.edu.pl/347/

Relevant publications 2/2

Bouda, Peter and Michael Cysouw (2012). *Treating Dictionaries as a Linked-Data Corpus*. In: Linked Data in Linguistics. Springer. http://link.springer.com/chapter/10.1007%2F978-3-642-28249-2_2

Yann Le Cun, Léon Bottou, Andrei Erofeev, Patrick Haffner and Bill W. Riemers (2001). "DjVu document browsing with on-demand loading and rendering of image components", *Internet Imaging*, San Jose.
http://leon.bottou.org/papers/lecun-2001

More resources available from the Digital Library at
http://bc.klf.uw.edu.pl/

Thank you for your attention!

jsbien@uw.edu.pl j.bilinska@uw.edu.pl m.sarnecki@uw.edu.pl

KLF UW

Department of Formal Linguistics
University of Warsaw