# Problémy a výsledky počítačového spracovania diela *Slowár Slowenskí Češko-Laťinsko-Ňemecko-Uherskí seu Lexicon Slavicum Bohemico-Latino-Germanico-Ungaricum*

*Radovan Garabík – Michaela Kajanová*

## 1. Úvod

*Slowár Slowenskí Češko-Laťinsko-Ňemecko-Uherskí seu Lexicon Slavicum Bohemico-Latino-Germanico-Ungaricum* (Bernolák, 1825; ďalej Bernolákov slovník), v modernej slovenčine *Slovník slovenský česko-latinsko-nemecko-maďarský alebo slovník slovenský česko-latinsko-nemecko-maďarský*, je prvý veľký prekladový slovník slovenčiny, ktorý vyšiel až po autorovej smrti v roku 1825 zásluhou J. Palkoviča. Slovníkom sa Bernolák snažil zavŕšiť svoje úsilie kodifikovať spisovnú slovenčinu. Za základ si zvolil jazyk vzdelancov západoslovenských kultúrnych centier, ako je Trnava a Bratislava.

Slovník je prvým veľkým prekladovým slovníkom slovenčiny, pozostáva z piatich dielov, v ktorých sú heslá uvádzané spôsobom: slovenské slovo – latinské ekvivalenty, nemecké ekvivalenty, maďarské ekvivalenty, české ekvivalenty[1], slovenské synonymá, a zo 6. dielu (tzv. Repertorium), ktorý obsahuje index latinských, nemeckých a maďarských slov s odkazmi na príslušné slovenské heslá. Pozostáva z 88 508 hesiel na 5 273 stranách (4 417 strán má prvých 5 zväzkov a 856 strán šiesty zväzok).

Slovník nie je chránený autorským zákonom, lebo A. Bernolák nebol občanom Slovenskej republiky a ani jeho dielo nebolo publikované v Slovenskej republike (a teda sa na neho nevzťahuje slovenský autorský zákon). Navyše, jeho dielo už nie je chránené ani v USA (jedna z krajín s najdlhšou dobou autorskoprávnej ochrany – 100 rokov po smrti autora), takže jeho digitalizácii, ďalšiemu spracovaniu a sprístupneniu verejnosti nestoja v ceste žiadne legálne prekážky.

## 2. Prvý krok: scan

Prvý krok k digitalizácii Bernolákovho slovníka sme urobili[2] už pred niekoľkými rokmi v rámci sprístupňovania najvýznamnejších slovenských jazykových zdrojov na webovej stránke Jazykovedného ústavu Ľ. Štúra SAV (ďalej JÚĽŠ). Scan slovníka bol vyhotovený s rozlíšením 600 DPI v 24-bitovej farebnej hĺbke a na archivačné účely je uložený v bezstratovom PNG formáte (ISO/IEC 15948:2004). Na sprístupnenie verejnosti sme použili formát DJVU (Bottou – Haffner – Howard – Simard – Bengio – LeCun, 1998). Ide o formát navrhnutý špeciálne na uchovávanie scanovaných dokumentov, ktorý ponúka vynikajúci kompresný pomer pre bitonálne obrazy (stratovú aj bezstratovú kompresiu), a separátny formát založený na vlnkovej kompresii (IW44), určený pre plnofarebné fotografické obrázky. DJVU súbory majú aj možnosť vloženia textovej vrstvy (typicky využívanú pre OCR), avšak v nami zverejnenom slovníku sa táto vlastnosť nevyužila.

[1] Ak sa odlišujú od slovenského tvaru.

[2] http://www.juls.savba.sk/ediela/bernolak/

Na stránke sme zverejnili prvých 5 zväzkov slovníka, v tvare bitonálneho čiernobieleho DJVU v pôvodnom rozlíšení 600 DPI a v podobe farebného DJVU v zníženej kvalite a v redukovanom rozlíšení 300 DPI. Veľkosť DJVU dokumentov je v rozmedzí 50 až 100 megabajtov, asi o tretinu viac pre farebné obrázky. Pre používateľov, ktorí nedisponujú možnosťou prezerania si DJVU súborov, sme poskytli aj jednotlivé dvojstránky v PNG formáte so zníženou kvalitou. Základným nedostatkom takto zverejnených dokumentov je to, že v nich nie je možné nijako vyhľadávať, ani k nim neexistuje heslár.

**3. Druhý krok: dict**

Hlavným zmyslom počítačovo spracovaného slovníka je možnosť rýchlo vyhľadávať jednotlivé heslá, bez zdĺhavého listovania či už v papierovej podobe, alebo v mnohostranovom grafickom súbore. Na webovej stránke[3] slovníkov JÚĽŠ sú sprístupnené rôzne slovníky a iné diela, spracované vo formáte vhodnom pre dict protokol (Faith – Martin, 1997), prostredníctvom ktorého prebieha vyhľadávanie prezentované cez www rozhranie. Systém umožňuje vyhľadávať heslové slová podľa prefixov, sufixov, častí slov a (v prípade prístupu cez IPv6) aj podľa regulárnych výrazov.

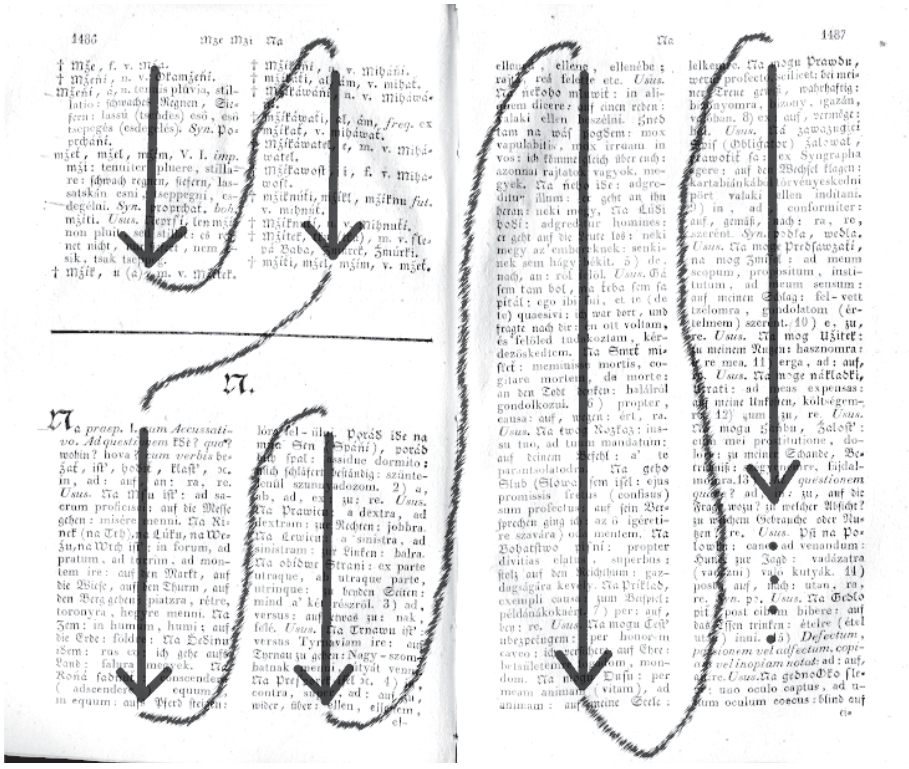
Aby sme umožnili vyhľadávanie v Bernolákovom slovníku podľa heslových slov, potrebujeme mať tieto heslové slová označené – OCR by bolo veľmi problematické, pretože všeobecne dobre fungujúci OCR systém pre švabach[4] neexistuje (a vyžadoval by si natrénovanie na Bernolákove znaky s diakritikou). Preto sme pristúpili k ručnému spracovaniu textu. Vzhľadom na náročnosť celého projektu sme zachovali samotné heslá iba ako obrázky, označili sme iba kľúčové slová.

Na vnútorné spracovanie a ukladanie dát (t. j. obrázok ako pozadie, s textovou vrstvou, kde by sa pridávali a editovali kľúčové slová) sme vybrali SVG formát (Eisenberg, 2002). Medzi jeho hlavné prednosti patrí pomerne jednoduché parsovanie a počítačové spracovanie (ide o formát založený na XML). Spracovanie prebiehalo v grafickom vektorovom editore InkScape (Bah, 2007). V záujme ergonómie spracovania a dosiahnutia čo najpriaznivejšieho času sme do súboru vkladali iba kľúčové slová umiestnené na rovnakom mieste, na ktorom sa nachádzali na obrázku, ktorý bol v pozadí.

Heslo (v grafickej podobe ako časť obrázka) je potom jednoznačne určené ako obrázok siahajúci od aktuálneho kľúčového slova po najbližšie nasledujúce slovo. Analýza výsledného označkovaného súboru je skomplikovaná tým, že typograficky je Bernolákov slovník vysádzaný v dvojstĺpcovom usporiadaní, heslo často presahuje do nasledujúceho stĺpca, niekedy aj na nasledujúcu stranu, a je veľké množstvo hesiel, ktoré svojím rozsahom presahujú niekoľko strán. Vyskytli sa však aj vážnejšie komplikácie – nie vždy je postupnosť textu lineárna. Na obr. 1 je príklad komplikovanej postupnosti textu pri začiatku nového písmena, obsahujúci šesťstranový výklad slova *na*. Rozdeľovanie takýchto úsekov sme vyriešili tabuľkou výnimiek pre konkrétne miesta, kde text nepokračoval priamo pozdĺž stĺpca (typicky na konci jedného písmena a začiatku ďalšieho).

[3] http://slovniky.korpus.sk/

[4] Zaujímavé je, že slovenský a český text je v Bernolákovom slovníku vysádzaný švabachom, ale nemecký text fraktúrou. Latinský a maďarský je vysádzaný antikvou.



Obr. 1: Príklad nelineárnej postupnosti textu v Bernolákovom slovníku

Rozdelenie dvojstránky na ľavú a pravú stranu, ako aj rozdelenie každej z týchto strán na dva stĺpce je robené iba na základe ich geometrickej pozície v grafickom súbore. Aby sme zamedzili prípadnému odrezaniu časti textu pri mierne posunutom obraze, pridali sme k okrajom jednotlivých stĺpcov niekoľko pixelov ako tzv. „bezpečnostnú zónu“.

### 3.1 Indexovanie heslových slov

Pri prepise používame pôvodný pravopis. Aj keď by mal prepis do súčasného pravopisu niekoľko výhod (hlavne možnosť vyhľadávať podľa súčasného, známeho tvaru slova), na druhej strane by znamenal stratu možnosti vyhľadávať podľa pôvodného pravopisu a nemožnosť vytvorenia pôvodného heslára. Taktiež by vznikli ťažko riešiteľné otázky, do akej miery prepisovať či prekladať slová, ktoré nemajú priamy ekvivalent v modernej slovenčine. Ako vhodný kompromis sme zvolili možnosť indexovať jednu heslovú stať viacerými kľúčovými slovami, pričom jedno z nich je v pôvodnom pravopise a prípadné ďalšie vzniknú transformáciou podľa niekoľkých jednoduchých automatických pravidiel, uvedených v tab. 1. Transformácie prebiehajú v poradí, v akom sú uvedené. Toto umožňuje získať príslušnú heslovú stať pre veľké množstvo vyhľadávaných slov aj v prípade zadania hľadaného výrazu v modernejšom pravopise. Je to doplnková funkcia pre tých používateľov, ktorí nemajú celkom zažitý originálny historický pravopis. Toto vyhľadávanie je potom doplnené možnosťou vyhľadávať slová s ignorovaním diakritiky (štandardne vo webovom rozhraní JÚĽŠ k slovníkom).

| pôvodný reťazec | g | ğ | w | ů | ňi | ňe | ďi | ďe | ťi | ťe | ňí$[5] | ďí | ťí | ňí |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| zmenený reťazec | j | g | v | ô | ni | ne | di | de | ti | te | nie | dí | tí | ní |

Tab. 1: Transformácie heslových slov počas indexovania

Pri spracovaní slovníka sme v heslových slovách našli niekoľko nečakaných javov:

- písmeno *ľ* sa v heslári vyskytuje 2 851-krát,
- apostrof sa vyskytuje 2 936-krát (často namiesto mäkčeňa, ale aj v niektorých iných prípadoch),
- ypsilon sa vyskytuje 99-krát (zväčša nahrádza písmeno *j*),
- *ě* (*e* s mäkčeňom) sa vyskytuje 33-krát,
- *'e* (apostrof, za ktorým nasleduje písmeno *e*, pravdepodobne omylom použitý namiesto chybne použitého *ě*) sa vyskytuje 987-krát,
- *e'* (*e*, za ktorým nasleduje apostrof, pravdepodobne omylom použité namiesto *'e*, ktoré je chybne používané namiesto omylom použitého *ě*) sa vyskytuje 2-krát.

Pri spracovaní sme tiež našli väčšie množstvo chýb alebo preklepov. Objavili sa prípady, keď chýbajú mäkčene (*grobianski, Gróf, Grófik* namiesto *ğrobianski, Ğróf, Ğrófik*), keďže samotné *g* bez mäkčeňa má význam dnešného *j* a hláska */g/* bola v Bernolákovom slovníku zapisovaná ako *ğ*. Tiež sú pri niektorých prípadoch zamenené písmená: *eňiptčanow* namiesto *eğiptčanow*, *Gegnodušnosť* namiesto *Gednodušnosť*. Nachádzame aj slová s vynechaným písmenom: *pašliwe* namiesto *plašliwe*, prípadne s písmenom navyše: *občerweňení* namiesto *očerweňení*. Apostrof namiesto mäkčeňa je napr. vo výrazoch: *nepow'edom'e* namiesto *nepowědomě*, *Neskrownosk'* namiesto *Neskrownosť* (v tomto slove je aj koncové *t* nahradené písmenom *k*), *Krest'an* namiesto *Kresťan*.

Do elektronickej podoby sme tieto slová prepísali presne tak, ako sa nachádzajú v slovníku, avšak pri vyhľadávaní ich nie je problém nájsť ani pod správnou podobou, pretože rozhranie ponúka pri vyhľadávaní neexistujúceho slova zoznam podobných slov odlišujúcich sa od hľadaného slova o Levenštejnovu vzdialenosť 1 (Левенштейн, 1965).

### 3.2 Spracovanie označkovaných dát

Po označkovaní heslových slov v SVG súboroch sme pristúpili k ich spracovaniu. Keďže nás zaujímajú súradnice heslových slov, z pomerne komplikovanej XML štruktúry sme vybrali iba atribúty *x* a *y* tagu *rect* v rámci jedného *flowRegion*. Vybraný obrázok zodpovedajúci jednému heslovému slovu bude potom všetko až po nasledujúce súradnice v nasledujúcom *flowRegion*. Na ilustráciu uvádzame obr. 2, ktorý prezentuje príklad vnútornej štruktúry SVG súboru.

[5] Regulárny výraz – t. j. skupina písmen *ňí* na konci slova je nahradená reťazcom *nie*. Toto zachytí časté prípady deverbatív.

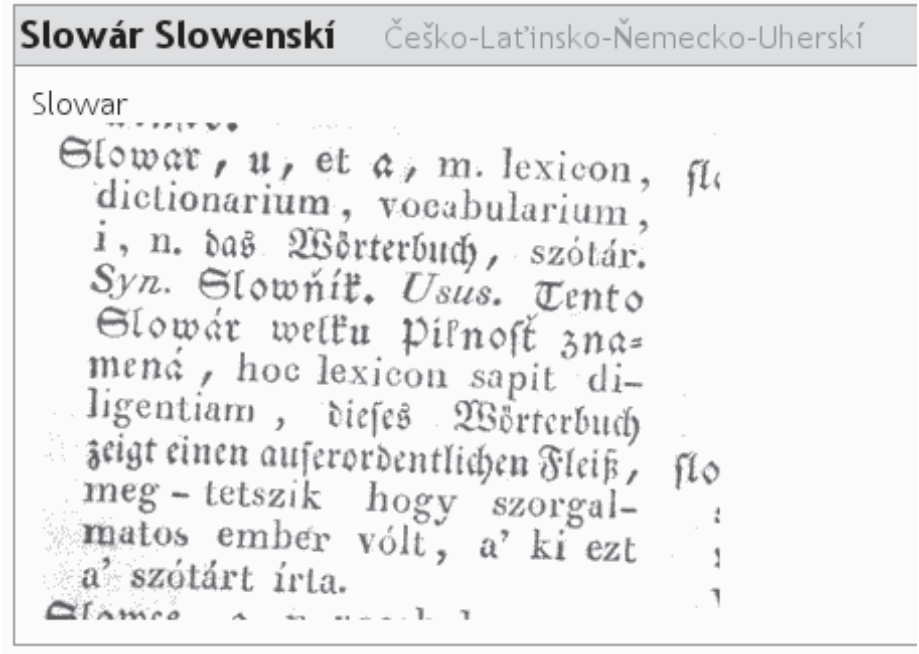
```
  <g
     inkscape:label="Vrstva 1"
     inkscape:groupmode="layer"
     id="layer1">
    <image
       sodipodi:absref="/home/extras/extra10/bernolak/2/scan0004.png"
       xlink:href="scan0004.png"
       width="2861"
       height="2465"
       id="image2471"
       x="-870.5"
       y="-342.995" />
    <flowRoot
       xml:space="preserve"
       id="flowRoot2475"
       style="font-size:40px;font-style:normal;font-weight:normal;fill:#000000;fill-opacity:1
;stroke:none;stroke-width:1px;stroke-linecap:butt;stroke-linejoin:miter;stroke-opacity:1;font
-family:Bitstream Vera Sans"><flowRegion
         id="flowRegion2477"><rect
           id="rect2479"
           width="241.42857"
           height="41.42857"
           x="-787.85712"
           y="-161.92354" /></flowRegion><flowPara
         id="flowPara2481">Kachowitost</flowPara></flowRoot>    <flowRoot
       xml:space="preserve"
       id="flowRoot2483"
       style="font-size:40px;font-style:normal;font-weight:normal;fill:#000000;fill-opacity:1
;stroke:none;stroke-width:1px;stroke-linecap:butt;stroke-linejoin:miter;stroke-opacity:1;font
-family:Bitstream Vera Sans"><flowRegion
         id="flowRegion2485"><rect
           id="rect2487"
           width="167.85715"
           height="42.857143"
           x="-787.85712"
           y="-119.78068" /></flowRegion><flowPara
         id="flowPara2489">Kačica</flowPara></flowRoot>    <flowRoot
       xml:space="preserve"
```

Obr. 2: Príklad vnútornej štruktúry SVG súboru

Na grafické spracovanie používame program *imagemagick* (Still, 2005), ktorý slúži na vyrezanie konkrétneho obdĺžnika zodpovedajúceho danému heslu, ale aj na úpravu obrázka na prezentačné účely. Ako najvhodnejšie sa ukázalo zmenšiť rozlíšenie na polovicu, mierne upraviť úroveň jasu – stmaviť tmavšie časti (t. j. písmo), zosvetliť svetlé (t. j. pozadie) a znížiť počet farieb na 15. Príkaz upravujúci jednotlivé obrázky potom vyzerá takto:

convert vstup.png -level 10,85% -auto-level -resize 50% -colors 15 +dither vystup.png

Jednotlivé časti hesla sú potom uložené v PNG formáte, pričom ako meno súboru je použité priamo heslové slovo. V prípade komplikovanejších hesiel presahujúcich hranice jedného stĺpca sú k názvu pridané arabské číslice. V slovníkovom dict súbore sú uložené iba mená týchto súborov, zobrazenie príslušných obrázkov zabezpečuje www rozhranie. Príklad zobrazeného hesla je uvedený na obr. 3.




Obr. 3: Ukážka vyhľadávania slova *Slowar*. V hornej, dolnej a pravej časti obrázka vidieť presah predchádzajúceho a nasledujúceho hesla a susedného stĺpca. Heslové slovo je uvedené s krátkym *a*, avšak vo výklade hesla, ako aj v názve slovníka sa už vyskytuje v podobe *Slowár*.

**4. Záver**

Digitalizácia Bernolákovho slovníka sleduje niekoľko cieľov. Jednak je to spôsob uchovania cenného historického diela, ktorý nepodlieha klasickej časovej degradácii (ako je žltnutie a rozpadanie papiera, blednutie tlače a pod.), jednak ponúka aj možnosť, ako sprístupniť toto dielo čitateľom, ktorí by sa k nemu ináč dostali len veľmi ťažko.

Indexované dielo s možnosťou vyhľadávania potom ponúka nové možnosti pre profesionálny, ale aj amatérsky výskum histórie slovenského jazyka. Slovník je prístupný (spolu s ostatnými slovníkmi a inými databázami JÚĽŠ) na adrese http://slovniky.korpus.sk/?d=bernolak.

## Literatúra

ЛЕВЕНШТЕЙН, Владимир Иосифович: Двоичные коды с исправлением выпадений, вставок и замещений символов. Докл. АН СССР 1965, выпуск 163, номер 4, с. 845 – 848.

BAH, Tavmjong: Inkscape: Guide to a Vector Drawing Program. Upper Saddle River: Prentice Hall Press 2007. ISBN 978-0137051731.

BERNOLÁK, Antoninus: Slowár Slowenskí Češko-Laťinsko-Ňemecko-Uherskí seu Lexicon Slavicum Bohemico-Latino-Germanico-Ungaricum. Budae: Typis et Sumtibus Typogr. Reg. Univers. Hungaricae 1825.

BOTTOU, Léon – HAFFNER, Patrick – HOWARD, Paul G. – SIMARD, Patrice – BENGIO, Yoshua – LECUN, Yann: High quality document image compression with "DjVu". J. Electron. Imaging 7, 410 (1998).

EISENBERG, J. David: SVG Essentials. Sebastopol: O'Reilly & Associates, Inc. 2002. ISBN 0-596-00223-8.

FAITH, Rik – MARTIN, Bret: "A Dictionary Server Protocol". [online]. Request for Comments 2229, Network Working Group 1997. Dostupné na: <http://tools.ietf.org/html/rfc2229>.

Information technology – Computer graphics and image processing – Portable Network Graphics (PNG): Functional specification (ISO/IEC 15948:2004) 2004.

STILL, Michael: The Definitive Guide to ImageMagick. Apress 2005. ISBN 978-1-59059-590-9.

**Problems and Results of Computer Processing of *Slowár Slowenski Češko-Laťinsko-Ňemecko-Uherski seu Lexicon Slavicum Bohemico-Latino-Germanico-Ungaricum***

The article describes the digitalization of the first large (pentalingual) Slovak dictionary written by A. Bernolák and published posthumously in 1825. The dictionary has been scanned and released on the webpages of the Ľ. Štúr Institute of Linguistics of the Slovak Academy of Sciences in DJVU and PNG formats. The keywords have been manually transcribed and the dictionary has been indexed for use by the dictd program, enabling queries via the dictionary web interface. The query interface offers several search strategies (exact, prefix, suffix, substring and regular expression) and returns the results in graphical format as an image of the dictionary entry and is available at the URL http://slovniky.korpus.sk/?d=bernolak.

RNDr. Radovan Garabík
Jazykovedný ústav Ľudovíta Štúra SAV, Bratislava, Slovenská republika
e-mail: garabik@kassiopeia.juls.savba.sk

Bc. Michaela Kajanová
Filozofická fakulta, Trnavská univerzita v Trnave, Slovenská republika
e-mail: michaela.kajanova@gmail.com