

# Analýza staročeské morfologie v Excelu



**BORIS LEHEČKA, [BORIS@DALIBORIS.CZ](mailto:BORIS@DALIBORIS.CZ)**  
**ODDĚLENÍ VÝVOJE JAZYKA**  
**ÚSTAV PRO JAZYK ČESKÝ AV ČR**

**LINGVISTIKA PRAHA 2014**  
**11. DUBNA**  
**16.00**

# Boris Lehečka



- [boris@daliboris.cz](mailto:boris@daliboris.cz)
- oddělení vývoje jazyka ÚJČ AV ČR
- programátor s lingvistickými základy
- Vokabulář webový <<http://vokabular.ujc.cas.cz>>
- materiály ke stažení
  - <http://vokabular.ujc.cas.cz/informace.aspx?t=LP2014>
  - <http://bit.ly/1qmOjzD>

# Obsah



- **Účel analýzy**
  - deklinace staročeských apelativ
  - časové období
  - vzory
- **Auditorium**
  - anketa
- **Excel**
  - Power Query
    - ✦ import a transformace datových zdrojů
    - ✦ programovací jazyk

# Obsah



- Excel
  - PowerPivot
    - ✦ stamiliony položek
    - ✦ tabulky
    - ✦ relace
  - Kontingenční tabulky a grafy
- Vstupy
  - výchozí
  - po transformaci
    - ✦ pomůcky
- Ukázka

# Účel analýzy



- deklinace staročeských apelativ
  - disertační práce Pavlína Jínové
- časové období
- vzory
  - lemmata
  - koncovky
- ověření výskytu tvaru/tvarů

# Anketa



- Kdo používá Excel?
- Jakou verzi Excelu?
  - 2010
  - 2013
  - jinou (např. Office 365 pro vysokoškoláky)
- Jakou edici Excelu?
  - Home and Student
  - Professional
- Kdo zná PowerPivot?
- Kdo zná Power Query?
- Kdo zná vertikálu?
- Příprava dat, nebo kontingenční tabulky/grafy v Excelu?

# Vstupy



- **Staročeská textová banka**
  - **Metadata**
    - ✦ identifikátor
    - ✦ zkratka
    - ✦ období vzniku
    - ✦ literární žánr
    - ✦ atp. (podle potřeby)
  - **Vertikála**
    - ✦ poznámky = metainformace

# Staročeská textová banka – metadata

Správa elektronických přepisů textů - [\* Správa přepisovaných textů - 1]

Šoubor Úpravy Zobrazit Nástroje Okno nápověda

216 z 336 filtry: soubor, autor, titul Zobrazit

Poř.	Soubor	Autor	Titul	Signatura	Datace	Využití	Čas. zařazení	Zpracování	Změněno	Památka	Pramen	Lit. druh
212	LetX.doc		[Staré letopisy ...		2. polovin...		do roku 1500	exportováno	30.1.2011 20...	Let	LetX	próza
213	ListKazimir.doc	Kazimír Jage...	[List]		1455		do roku 1500	exportováno	4.5.2011 16:03	List	ListKazimir	próza
214	ListLaci.docx		[List láci]		1. polovin...		do roku 1500	exportovat	20.1.2014 11...	List	ListLaci	próza
215	ListLit.doc		[Připisek v zakl...		začátek 1...		do roku 1500	exportováno	20.5.2011 18...	ListLit	ListLit	próza
216	ListZikm.doc	Zikmund Luc...	[List krále Zikm...		1431		do roku 1500	exportováno	30.1.2011 20...	ListZikm	ListZikm	próza
217	Lucidar.doc		Lucidář		1498		do roku 1500	exportováno	30.1.2011 20...	Lucid	LucidT	próza
218	Lucidar1750....		Lucidář, totiž k...		1750-1775		do roku 1800	exportováno	30.1.2011 20...	Lucid		próza
219	LyraMat.docx	Mikuláš z Lyry	[Výklad evange...		začátek 1...		do roku 1500	exportovat	15.1.2012 14...	LyraMar	LyraMat	próza
220	LyrDuch.docx		[Nejstarší česk...		14. a 15. ...		do roku 1500	přepsáno	10.4.2010 15...	LyrDuch	LyrDuch	verš
221	LyrVil.docx		[Staročeská lyri...		14. a 15. ...		do roku 1500	přepsáno	10.4.2010 16...	LyrVil	LyrVil	verš
222	MeiCesA.docx		[MeiCesA.docx]		14. a 15. ...		do roku 1500	přepsáno	10.4.2010 16...	MeiCesA	MeiCesA	próza

Přehled Zpracování Pramen Edice Vokabulář Manuscriptorium

autor: Zikmund Lucemburský  uzaální titul: [List krále Zikmunda panu Haškovi]  uzaální

datace: 1431 tiskař: místo tisku:

století: 1400, polovina: 1, desetiletí: 3, rok: 1431, upřesnění: , relativní chronologie: 3 typ předlohy přepisu: rukopis foliace/paginace: 66rv

instituce: Národní knihovna České republiky město: Praha signatura: III G 16

země: Česko

ukázka:



# Staročeská textová banka – Word

66r Copia littere Sigismundi regis Ungarie pro domino Hasskone.

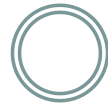
**Zikmund, z Božie milosti římský král, po vše časy rozmnožitel říše a uherský, český etc. král**

Urozený, věrný, milý, psal nám urozený Menhard z Hradce, kterak s zámkov syna našeho, kniežete rakúského, jsú jemu mnoho jeho lidí zjimali, niekoľiko vozov vina i koní pobrali, a že služebníci kniežetini ještě vždy na silnici lidem berú, a to vše v tom přimíří až do sv. Jiřie skrze nás dělanémpůvodně napsáno „dielanym“, poté stejnou rukou nadepsáno -e. A ač jsem syna našeho často obeslal, však, jakož nám píše, tehdy mu se ještě nic neodestálo. A dává nám Hradečský vinu, že to pro nás trpí etc. I psali sme synu našemu, aby neučinil jinak, než jemu ty vězně propustil a pobrané věci vrátiti kázal. A k tomu aby to přimířie s Hradeckým od sv. Jiřie až do roka protáhl a zjednal, aby držáno bylo. Těž sme také Hradeckému psali, aby to přimířie do roka s synem naším prodil až do roka. Protož, aby ty věci spieše k miestu přišly, žádámy od tebe, aby se to tak podlé žádanie našeho stalo, jakož sme pak oběma o to psali, že sme tobě to poručili. Pakli by která nesnáze veliká mezi nimi byla, ješto by tomu co činiti nemohl, tehdy vždy se toho snaž, aby to v dobrotě státi nechali až na nás a to přiměřie a ty věci aby před se šly.

Také věděti dáváme, že z Čech listy sme měli, kterak Pražené a strana odporná přichylny jsú k tomu svatému concilium do Bazle Bazel Bazel poslati a že jsú již na některých artykulech přestali, ač se někteří mezi nimi těm protivie. I slyšimy, že hlas v Čechách běží, že by to concilium nemělo před se jíti a že se lidé tiem rozpáčejí. Protož věz, že je otec náš svatý papež chtěl to concilium pól druhého léta do Bononije protáhnúti pro někaké příčiny a byl je bully na to vyslal. Ale to svaté concilium, my i jiní králi a kniežata křesťanská ustanovili sme se, že to vždy bez ponúcenie ponúcenie ponuzenie 66v jmá před se jíti, a na to sme k jeho svatosti poslali a tak že to před se pójde. Protož žádáme a prosíme, aby jim to zjevil a je zpravil a navedl podlé

Stránka: 1 z 2 Slova: 604 Čeština 110%

# Vstupy



- **Slovníky**
  - ESSČ (*Přib–ž, ž–ch*)
  - MSS (*a–ž*)
  - StčS (*n–při*)
  - GbSlov (*a–netbanlivý*)
  - formát XML

# Úpravy vstupů



- **Vertikála**
  - kategorizace tokenů
    - ✦ jazyk
    - ✦ torzo
    - ✦ funkce (interpunkce, číslo)
    - ✦ relevance
  - segmentace tokenů na fonogramy
    - ✦ zakončení
      - 1–3 fonogramy

# Fonogram



- grafická jednotka korespondující s fonémem
  - změna fonému v důsledku hláskoslovného vývoje, flexe a/nebo slovo tvorby znamená změnu fonogramu

<b>m</b>	<b>ú</b>	<b>ch</b>	<b>a</b>
<b>m</b>	<b>ú</b>	<b>š</b>	<b>ě</b>
<b>m</b>	<b>ou</b>	<b>š</b>	<b>e</b>

<b>d</b>	<b>ie</b>	<b>v</b>	<b>k</b>	<b>a</b>
<b>d</b>	<b>í</b>	<b>v</b>	<b>k</b>	<b>a</b>

<b>h</b>	<b>o</b>	<b>s</b>	<b>t</b>	<b>ie</b>
<b>h</b>	<b>o</b>	<b>s</b>	<b>t</b>	<b>í</b>

<b>h</b>	<b>o</b>	<b>s</b>	<b>t</b>	<b>i</b>	<b>e</b>
<b>h</b>	<b>o</b>	<b>s</b>	<b>t</b>	<b>i</b>	<b>í</b>

# Úpravy vstupů



- **Metadata o textech**
  - výběr údajů
  - datace
    - ✦ časová období
  - převod z XML na CSV
- **Slovníky**
  - lemma
  - slovní druh
  - morfologická charakteristika

# Úpravy vstupů – nástroje



- C#
  - transformace
    - ✦ DOCX > XML
    - ✦ XML > VERT
      - VERT > TAB
    - ✦ Metadata > TAB
- OpenRefine
  - <http://openrefine.org>
  - analýza vstupů
    - ✦ statistiky
    - ✦ redefinice vlastností

# Excel – Power Query



- **Doplněk**
  - verze 2.10.3598.81
  - pro Excel 2010 a 2013
  - podmínky
    - ✦ Windows Vista až Window 8.1
    - ✦ Office 2010 Professional Plus + SA
    - ✦ Office 2013 Professional Plus, Office 365 ProPlus, Excel 2013
- **Princip**
  - import dat z různých zdrojů
  - nový programovací jazyk
  - není třeba řešit zabezpečení maker

# Excel – PowerPivot



- **Microsoft SQL Server 2012 PowerPivot for Excel**
- **Doplněk**
  - pro Excel 2010
  - od Excelu 2013 je již součástí programu
  - verze 11.0.3129.0
- **Podmínky**
  - Windows XP SP3 až Window 8
  - Office 2010 zdarma pro všechny edice
  - v Excelu 2013 návrh pouze v edici Professional
- **64bitová verze (pro velké objemy dat)**

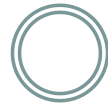


# Excel – PowerPivot



- Princip
  - stamiliony záznamů
  - tabulky
  - relace
- Nevýhody
  - nerozlišuje velikost písmen
  - relace
    - ✦ M : N
    - ✦ text != jedinečný klíč
  - textová data
    - ✦ velký objem
    - ✦ bez redukce

# Excel – PowerPivot



- Kontingenční tabulky
- Kontingenční grafy
- Průřezy

# Excel – doporučení



- **Vstupní data**
  - textové soubory
    - ✦ rychlé generování
    - ✦ lze i v Excelu
    - ✦ identifikátory
      - relace
  - tabulky v sešitě
    - ✦ ad hoc metadata
  - aktualizace
    - ✦ v PowerPivotu
    - ✦ v Excelu
- **Import v PowerPivotu**
  - nejprve nastavit parametry, pak vybrat soubor
    - ✦ první řádek jako názvy, až po výběru souboru

# Ukázka



- **Metadata**
  - XML
  - TXT
  - úpravy
- **Textová banka**
  - Word
  - XML TEI P5
  - Vertikála

# Ukázka



- **PowerPivot**
  - ✦ import dat
  - ✦ vytvoření relací
  - ✦ počítané sloupce
  - ✦ míry
  - ✦ testovací data
    - malý objem
- kontingenční tabulka
  - ✦ vytváření
  - ✦ interaktivita
  - ✦ průřezy
    - pro více objektů
- kontingenční graf
  - ✦ vytváření
  - ✦ interaktivita
  - ✦ průřezy

# Ukázka



- **Power Query**
  - vytvoření dotazu
  - opětovné spuštění dotazu
- **Sešit Tokeny**
  - reálná data
  - připravené tabulky
  - připravené grafy
  - tabulka a graf na přání

# Analýza morfologie v Excelu



- **Plusy**

- off-line
- není třeba korpusový manažer
- opakovatelnost (s jinými daty)
- ad hoc analýzy
- kontingenční přehledy
- zdarma (pro Excel 2010)

- **Minusy**

- příprava vstupních dat
- chybí kontext
- bez pokročilých analýz

# Postupy



- **Míry**
  - Formátování čísel
  - Velikost písmen
- **Identifikátory**
- **Aktualizace**
  - PowerPivot
  - Kontingenční tabulky
  - Propojené tabulky
  - Power Query
  - Soubory
- **Průřezy**